One Ecosystem 1: e8621
doi: 10.3897/oneeco.1.e8621



Review Article

# Adoption of Machine Learning Techniques in Ecology and Earth Science

Anne E Thessen ‡,§

‡ The Ronin Institute for Independent Scholarship, Montclair, United States of America
§ The Data Detektiv, Waltham, United States of America

Corresponding author: Anne E Thessen (annethessen@gmail.com)

Academic editor: Benjamin Burkhard

Received: 24 Mar 2016 | Accepted: 20 Jun 2016 | Published: 27 Jun 2016

Citation: Thessen A (2016) Adoption of Machine Learning Techniques in Ecology and Earth Science. One Ecosystem 1: e8621. doi: 10.3897/oneeco.1.e8621

## Abstract

### Background

The natural sciences, such as ecology and earth science, study complex interactions between biotic and abiotic systems in order to understand and make predictions. Machine-learning-based methods have an advantage over traditional statistical methods in studying these systems because the former do not impose unrealistic assumptions (such as linearity), are capable of inferring missing data, and can reduce long-term expert annotation burden. Thus, a wider adoption of machine learning methods in ecology and earth science has the potential to greatly accelerate the pace and quality of science. Despite these advantages, the full potential of machine learning techniques in ecology and earth science has not be fully realized.

### New information

This is largely due to 1) a lack of communication and collaboration between the machine learning research community and natural scientists, 2) a lack of communication about successful applications of machine learning in the natural sciences, 3) difficulty in validating machine learning models, and 4) the absence of machine learning techniques in a natural

science education. These impediments can be overcome through financial support for collaborative work and the development of graduate-level educational materials about machine learning. Natural scientists who have not yet used machine learning methods can be introduced to these techniques through Random Forest, a method that is easy to implement and performs well. This manuscript will 1) briefly describe several popular machine learning tasks and techniques and their application to ecology and earth science, 2) discuss the limitations of machine learning, 3) discuss why ML methods are underutilized in natural science, and 4) propose solutions for barriers preventing wider ML adoption.

## Keywords

ecology, machine learning, earth science, statistical learning

## Introduction

Machine Learning (ML) is a discipline of computer science that develops dynamic algorithms capable of data-driven decisions, in contrast to models that follow static programming instructions. The very first mention of 'machine learning' in the literature occurred in 1930 and use of the term has been growing steadily since 1980 (Fig. 1). While discussion of ML is likely to recall scenes from popular science-fiction books and movies, there are many practical applications of ML in a wide variety of disciplines from medicine to finance. Part of what makes ML so broadly applicable is the diversity of ML algorithms capable of performing very well under messy, real-world conditions. Despite, and perhaps because of this versatility, uptake of ML applications have lagged behind traditional statistical techniques in the natural sciences.

The advantage of ML over traditional statistical techniques, especially in earth science and ecology, is the ability to model highly dimensional and non-linear data with complex interactions and missing values (De'ath and Fabricius 2000, Recknagel 2001, Olden et al. 2008, Haupt et al. 2009b, Knudby et al. 2010a). Ecological data specifically are known to be non-linear and highly dimensional with intense interaction effects; yet, methods that assume linearity and are unable to cope with interaction effects are still being used (Olden et al. 2008, Knudby et al. 2010a) with some modification of the data to try and make the methods work (Knudby et al. 2010a, Pasini 2009, Džeroski 2001). Several comparative studies have already shown that ML techniques can outperform traditional statistical approaches in a wide variety of problems in earth science and ecology (Lek et al. 1996b, Levine et al. 1996, Manel et al. 1999, Segurado and Araújo 2004, Elith et al. 2006, Lawler et al. 2006, Prasad et al. 2006, Cutler et al. 2007, Olden et al. 2008, Zhao et al. 2011, Bhattacharya 2013); however, directly comparing the results of statistical techniques to ML techniques can be difficult and requires careful consideration (Fielding 2007).

The exact division between ML methods and traditional statistical techniques is not always clear and ML methods are not always better than traditional statistics. For example, a system may not be linear, but a linear approximation of that system may still yield the best predictor. The exact method(s) must be chosen based on the problem at hand. A meta approach that considers the results of multiple algorithms may be best. This manuscript will discuss four types of ML tasks and seven important limitations of ML methods. These tasks and limitations will be related to six types of ML techniques and their relative strengths and weaknesses in ecology and earth science will be discussed. Specific applications of ML in ecology and earth science will be briefly reviewed with the reasons ML methods are underutilized in natural sciences. Potential solutions will be proposed.

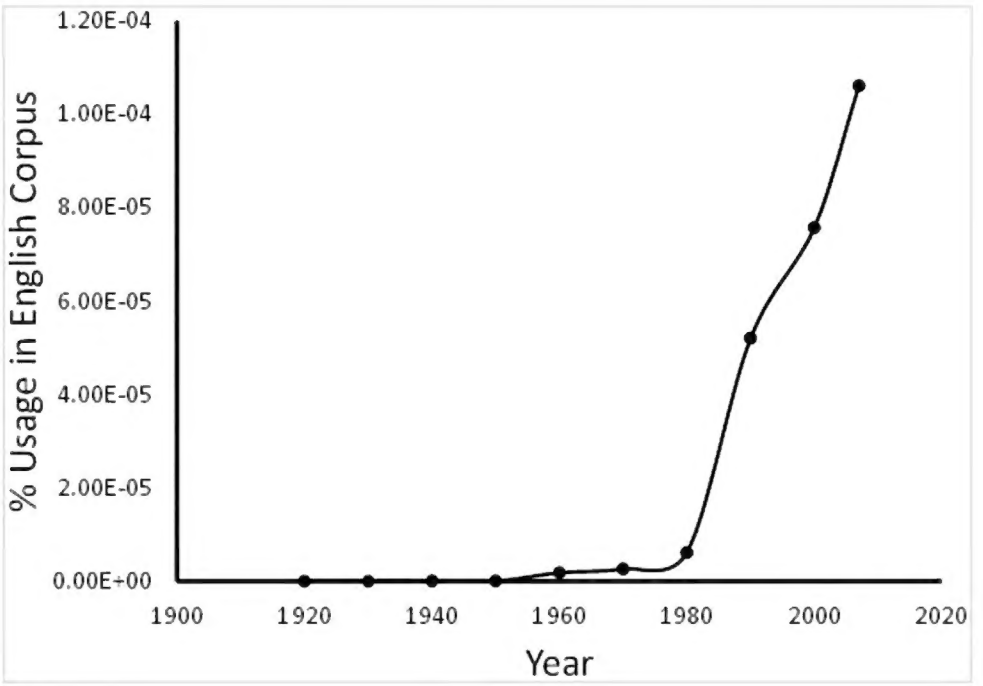


Figure 1.

**Use of the phrase 'machine learning' in the Google Books Ngram Viewer**: This plot shows the use of the phrase 'machine learning' by decade as percentage of total words in the Google English Corpus. http://books.google.com/ngrams

## Background

The basic premise of ML is that a machine (i.e., algorithm or model) is able to make new predictions based on data. The basic technique behind all ML methods is an iterative combination of statistics and error minimization or reward maximization, applied and combined in varying degrees. Many ML algorithms iteratively check all or a very high number of possible outcomes to find the best result, with "best" defined by the user for the problem at hand. The potentially high number of iterations is prohibitive of manual calculations and is a large part of why these methods are only now widely available to individual researchers.

Computing power has increased such that ML methods can be implemented with a desktop or even a laptop. Before the current availability of computing power, ecologists and earth scientists had to settle for statistical methods that assumed linearity (Knudby et al.

2010a) and limited, controlled experiments (Fielding 1999b). Both of these restrictions limit the scale of studies and accuracy of results. A similar acceleration has been observed for numerical modeling of natural systems, where model predictions have improved because increased computing power has allowed for the inclusion of more parameters and, more importantly, finer granularity (see Semtner 1995, Forget et al. 2015 for examples in oceanography).

The first step in applying ML is teaching the algorithm using a training data set. The training data set is a collection of independent variables with the corresponding dependent variables. The machine uses the training data to "learn" how the independent variables (input) relate to the dependent variable (output). Later, when the algorithm is applied to new input data, it can apply that relationship and return a prediction. After the algorithm is trained, it needs to be tested to get a measure of how well it can make predictions from new data. This requires another data set with independent and dependent variables, but the dependent variables (target) are not provided to the learner. The algorithm predictions (output) are compared to the withheld data (target) to determine the quality of the predictions and thus the utility of the algorithm. This comparison is an important difference between ML and traditional statistical techniques that use p values for validation.

> ***A Note on Terms***: The interdisciplinary nature of ML and its application has resulted in a confusing collection of terms for similar concepts. Below are groups of functional synonyms describing the major concepts discussed in this manuscript.
>
> - Observation, instance, data point: These terms are used to describe the data instances that can be thought of as rows in a spreadsheet.
> - Explanatory variables, features, input, independent variables, x, regressors: These terms are used to describe the independent variables/input data that are used to make predictions.
> - Outcomes, dependent variables, y, classes, output: These terms are used to describe the dependent variables/output that are the results of the algorithm or part of the training/test set. The outcomes in the test set that the algorithm is trying to predict are referred to as the "target".
> - Outlier, novelty, deviation, exception, rare event, anomaly: These terms are used to describe data instances that are not well represented in the data set. They can be errors or true outliers.

## Machine Learning Tasks

There are four different types of tasks that will be discussed in the context of available ML techniques and natural science problems. Most ML techniques can be used to perform multiple tasks and several tasks can be used in combination to address the same problem; therefore, it can be difficult to draw firm boundaries around categories of tasks and techniques. Many of the natural science problems discussed in the latter half of this paper have been addressed using all of the tasks and techniques discussed. The list below is not meant to be comprehensive. Only the tasks most relevant to the natural science

applications are discussed here. Each ML technique mentioned here is more thoroughly discussed in its own section.

Task 1) Function Approximation. In this task, the machine inferrs a function from (x,y) data points, which are real numbers (Bishop 2006, Alpaydin 2014). Regression and curve-fitting are types of function approximation. Artificial Neural Networks are one ML technique that performs function approximation (see discussion of ANN below). Natural science problems such as predicting the global riverine fish population (Guégan et al. 1998) and forecasting oceanographic conditions (Hsieh 2009) have been addressed with Artificial Neural Networks performing function approximation tasks. Tree-based methods can also be used for function approximation via regression (Loh 2014). Linear regression is an example of a traditional statistical method that performs a function approximation task (Sokal and Rohlf 2011).

Task 2) Classification. This process assigns a new observation to a category based on training data (Alpaydin 2014, Kotsiantis 2007). A common example of classification is the automated sorting of spam and non-spam email. ML techniques that are known to be good classifiers include Random Forest, Support Vector Machines, and Bayesian Classifiers, which will also output the probability that the observation belongs to the inferred class. Classification tasks in the natural sciences include the automated identification of species using recorded echolocation calls (Armitage and Ober 2010) and monitoring river water quality (Walley and Džeroski 1996), which have been performed using a Random Forest and a Bayesian Classifier, respectively. A linear discriminant analysis is an example of a traditional statistical method that can perform a classification task (Sokal and Rohlf 2011).

Task 3) Clustering. This task is similar to classification, but the machine is not given training data to learn what the classes are (Jain 2010). It is expected to infer the classes from the data. This task clusters data into groups such that objects in the same group are more similar than objects between groups. Each cluster is then an inferred class. Clusters have a situation-specific definition, thus there are several different clustering strategies available (e.g. hierarchical clustering, centroid clustering, etc). This task is often used for data exploration and knowledge discovery before another ML technique is applied. The Support Vector Machine and Artifical Neural Network (in the form of a Self Organizing Map) are two types of ML techniques that can perform a clustering task (Du 2010, Ben-Hur et al. 2001). Clustering can be used in the natural sciences to detect rare events (Omar et al. 2013), such as identifying a bird call in hours of streaming remote sensor data (Kasten et al. 2010).

Task 4) Rule Induction. This task extracts a set of formal rules from a set of observations, which can be used to make predictions about new data. Fuzzy Inference and some Tree-based ML techniques use rule induction to make predictions. Genetic Algoriothms can be used to infer rules (Fidelis et al. 2000, Sastry et al. 2013). Rule induction is a three-step process of 1) feature construction, where the features are turned into binary features, 2) rule construction, where features are searched for the combination that is most predictive for a class, and 3) hypothesis construction, where sets of individual rules are combined (Fürnkranz et al. 2012a) . Rule induction has been used in the natural sciences to predict

microbial biomass and enzyme activity in soil (Tscherko et al. 2007) and develop biodegradation models for industrial chemicals (Gamberger et al. 1996). A good example of overlap between categories of tasks and techniques is the decision tree, which uses rule induction to perform a classification task.

## Machine Learning Limitations

As with any technique, a working knowledge of the limitations of ML is necessary for proper application (Domingos 2012). Some limitations result from user misconceptions and some result from not recognizing problematic data. There are three major categories of mistakes that result from misconceptions of ML practitioners.

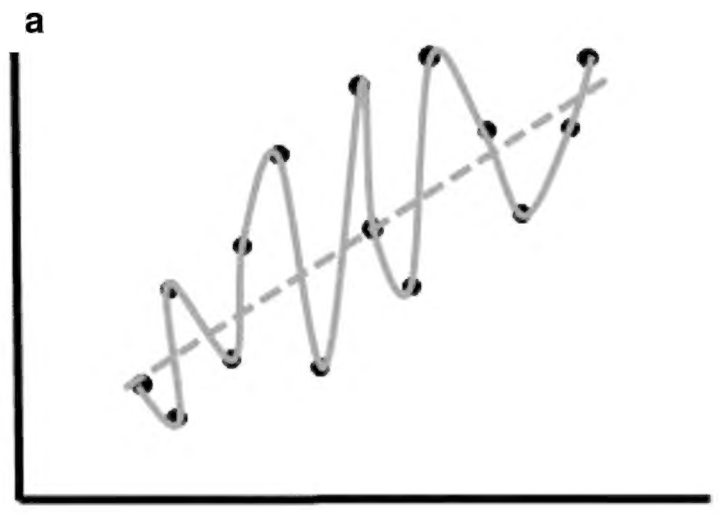


Figure 2.

Comparison of performance of two algorithms (grey lines) on hypothetical training (A) and test (B) data (black points).

**a**: Training Data. Algorithm 1 (solid line) models the training data perfectly, with no error. Algorithm 2 (dashed line) is much more generalized and does not model the training data as well as Algorithm 1.

**b**: Test Data. Algorithm 1 (solid line) modeled the training data perfectly, but has very high error on the test data. Algorithm 2 had a higher error on the training data, but models the test data with a reasonably low error. Algorithm 1 is an example of overfitting. Algorithm 2 is a much better real-world predictor.

1) Demanding Perfection. Algorithms that perfectly model training data are not very useful. This is due to ***overfitting*** and it happens when an algorithm is so good at modeling the training data that it does not perform well in the "real world" (Fig. 2Hawkins 2004). The prediction error given by the training data might be low, but the prediction error given by the test data, called the generalization error, is the measure of how well the algorithm will do in a real-world application. When a model is overfit, the prediction error is much lower on the training data than the test data. In general, as performance on the training data increases, performance on the test data will increase only to a point before decreasing (Fig. 3).


Figure 3.

As algorithm performance on training data increases, performance on test data increases only to a certain point (dashed line). Increases in performance on training data beyond this point results in overfitting.

2) Favoring Complexity Over Simplicity. It is important for a problem to be addressed with just the right amount of complexity and this varies according to the nature of the problem and the data (e.g. Merow et al. 2014). A more complex algorithm will not necessarily outperform a simpler algorithm (Olden et al. 2008, Domingos 2012). This misconception is related to overfitting because one way to improve model performance is to make it more complex, but that results in an ungeneralizable model. In many circumstances, more data with a simpler algorithm is better than a more complex algorithm (Domingos 2012). An iterative approach is often best, transitioning from simple to complex techniques and comparing the results.

3) Including As Many Features As Possible. It can be difficult to know *a priori* which features are important predictors in a given problem, but including a large number of features in a model (i.e. a shotgun approach), especially features that are not relevant, can make a model a poor predictor Keogh and Mueen 2011. This is because as the number of features increases, learners need a rapidly increasing amount of training data to become familiar with all combinations. As a result, algorithms with fewer features are often better predictors than algorithms using many features. This is often referred to as the ***Curse of***

***Dimensionality*** and can be addressed with feature selection or feature extraction (types of dimensionality reduction) before training the algorithm (Domingos 2012). Both of these methods reduce the number of features to a subset of the most important variables, either by identifying irrelevant and redundant features or by creating new, aggregate features.

The second type of limitation results from not recognizing imperfections in data sets (Kotsiantis et al. 2006). Problematic data are more the norm than the exception in real-world applications. Not recognizing and addressing this fact can cause serious complications. The following are common data problems.

1) Class imbalance. This problem occurs when one or more classes are underrepresented compared to the others (Japkowicz and Stephen 2002, Chawla 2005). With severe class imbalance, the distribution of the classes can vary broadly between the training and test sets, resulting in a non-generalizable model. There are no definitive rules about exactly when class imbalance is a problem or how to address it. A common solution is to balance the classes through upsampling or downsampling. Another coping strategy is to change the method of evaluation to more appropriately weight the correct inference of an underrepresented class (Japkowicz and Stephen 2002).

2) Too many categories. This problem is related to the Curse of Dimensionality and it occurs when a category has a high number of distinct values (i.e. high cardinality, Moeyersoms and Martens 2015). High-cardinality categories (such as zip code or bank account number) can be very informative, but can also increase the number of dimensions and thus decrease performance. One way to cope with a high-cardinality category is to combine levels using domain knowledge. For example, in the category "Taxon", instead of having a value for every species, have a single value for every Genus or higher rank. Another way to address high-cardinality is through data preprossessing and transformations that reduce the number of levels in the category (Micci-Barreca 2001, Moeyersoms and Martens 2015).

3) Missing data. Different types of learners and problems have different levels of tolerance for missing data during training, testing, and prediction (see discussion of ML techniques below). There are several methods for applying ML techniques to data with missing values (Saar-Tsechansky and Provost 2007, Gantayat et al. 2014). Techniques for coping with missing data include imputation, removal of the instance, or segmentation of the model (Saar-Tsechansky and Provost 2007, Gantayat et al. 2014, Jerez et al. 2010). The segmentation approach involves removing the features that correspond to the missing data (Gantayat et al. 2014). In some cases it may be worthwhile to acquire the missing data through more testing, sampling, or experimentation.

4) Outliers. If the observations are real and not the result of human error, outliers can be an important source of insight. They only become a problem when they go unnoticed and models are applied to a data set as though outliers are not present. There are many methods for outlier detection that are recommended as a preprocessing step before ML (Escalante 2005).

## Machine Learning Techniques

### Tree-based Methods

Tree-based ML methods include decision trees, classification trees, and regression trees (Olden et al. 2008, Hsieh 2009, Kampichler et al. 2010). For these methods, a tree is built by iteratively splitting the data set based on a rule that results in the divided groups being more homogeneous than the group before (Fig. 4). The rules used to split the tree are identified by an exhaustive search algorithm and give insight into the workings of the modeled system. A single decision tree can give vastly different results depending on the training data and typically has low predictive power (Iverson et al. 2004, Olden et al. 2008, Breiman 2001a). Thus, several ensemble-tree methods have been developed to improve predictive power by combining the results of multiple trees, including boosted trees and bagged trees (Breiman 1996, De'ath 2007). A boosted tree results from a pool of trees created by iteratively fitting new trees to minimize the residual errors of the existing pool (De'ath 2007). The final boosted tree is a linear combination of all the trees (Elith et al. 2008). Bagging is a method that builds multiple trees on subsamples of the training data (bootstrap with replacement) and then averages the predictions from each tree to get the bagged predictions (Breiman 1996, Knudby et al. 2010a).

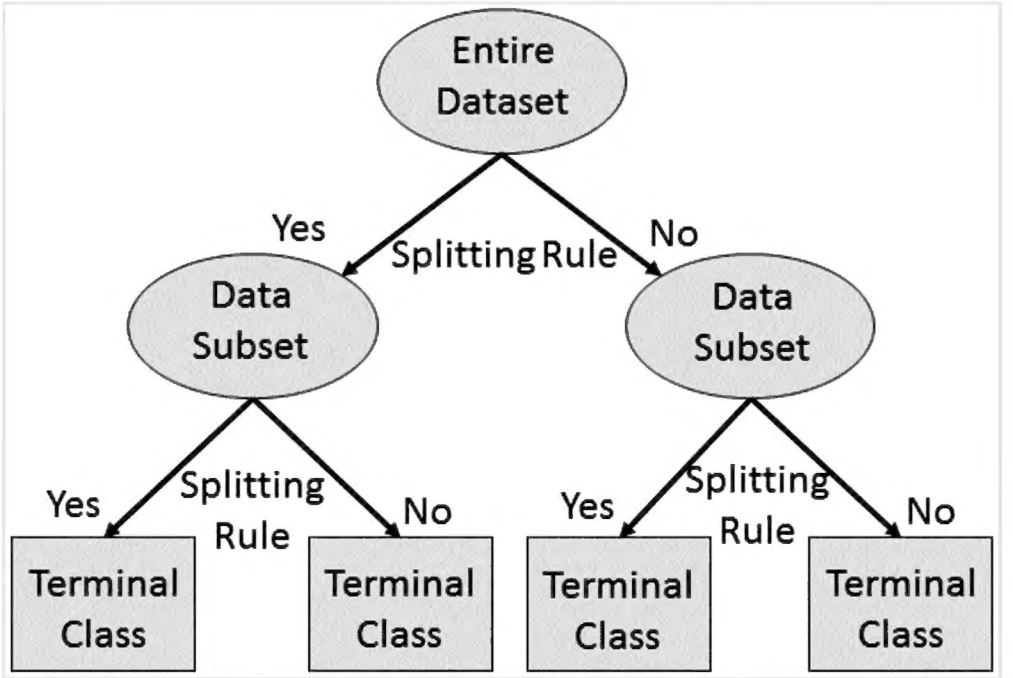


Figure 4.

**Decision and Classification Tree Schematic**: Tree-based machine learning methods infer rules for splitting a data set into more homogeneous data sets until a specified number of terminal classes or maximum variance within the terminal classes is reached. The inferred splitting rules can give additional information about the system being studied.

Random Forest is a relatively new tree-based method that fits a user-selected number of trees to a data set and then combines the predictions from all trees (Breiman 2001a). The Random Forest algorithm creates a tree for a subsample of the data set. At every decision only a randomly selected subset of variables are used for the partitioning. The predicted

class of an observation in the final tree is calculated by majority vote of the predictions for that observation in all trees with ties split randomly.

Ensemble tree-based methods, especially Random Forest, have been demonstrated to outperform traditional statistical methods and other ML methods in earth science and ecology applications (Cutler et al. 2007, Kampichler et al. 2010, Knudby et al. 2010a). They can cope with small sample sizes, mixed data types, and missing data (Cutler et al. 2007, Olden et al. 2008). The single-tree methods are fast to calculate and the results are easy to interpret (Kampichler et al. 2010), but they are susceptible to overfitting (Olden et al. 2008) and frequently require "pruning" of terminal nodes that do not give enough additional accuracy to justify the increased complexity (Breiman et al. 1984, Garzón et al. 2006, Cutler et al. 2007, Olden et al. 2008, Džeroski 2009). The ensemble-tree methods can be computationally expensive (Cutler et al. 2007, Olden et al. 2008, Džeroski 2009), but resist overfitting (Breiman 2001a). Random Forest algorithms can provide measures of relative variable importance and data point similarity that can be useful in other analyses (Cutler et al. 2007), but can be clouded by correlations between independent variables (Olden et al. 2008). Implementing Random Forest is relatively straightforward. Only a few, easy-to-understand parameters need to be provided by the user (Kampichler et al. 2010), but the final Random Forest does not have a simple representation that characterizes the whole function (Cutler et al. 2007). Tree methods also do not give probabilities for results, which means that data are classified into categories, but the probability that the classification is correct is not given.

For a more detailed description of tree-based methods see Breiman et al. 1984, Breiman 2001b, Loh 2014 and chapter 8 in James et al. 2013.

**Artificial Neural Networks**

An Artificial Neural Network (ANN) is a ML approach inspired by the way neurological systems process information (Recknagel 2001, Olden et al. 2008, Boddy and Morris 1999, Hsieh 2009). There are many types of ANNs, but only a few are typically used in earth science and ecology, such as the multi-layer, feed-forward neural network, which will be the focus of this section (Pineda 1987, Kohonen 1989, Chon et al. 1996, Recknagel 2001, Lek and Guégan 2000). A multi-layer ANN has three parts: 1) the input layer, which receives the independent variables 2) the output layer, where the results are found, and 3) the hidden layer, where the processing occurs (Fig. 5). Each layer is made up of several units (neurons). Each unit is connected to all the other units in the neighboring layer, but not the units in the same layer or in non-adjacent layers. Feed-forward ANNs allow data to flow in one direction, from input to output only. The number of units in the hidden layer can be changed by the user to optimize the trade-off between overfitting and variance (Camargo and Yoneyama 2001, Kon and Plaskota 2000. Too many units in this layer can lead to overfitting. Each connection between units has a weight. Training the ANN involves an iterative search for an optimal set of connection weights that produces an output with a small error relative to the target. After every iteration, the weights are adjusted to bring the output closer to the target using a back-propagation algorithm. Bayesian methods and Genetic Algorithms can also be used to find the optimal connection weights (Bishop 2006,

Kotsiantis et al. 2006, Siddique and Tokhi 2001, Yen and Lu 2000). Performance can be sensitive to initial connection weights, which are typically chosen randomly in the beginning, and the number of hidden units, so multiple networks should be processed while varying these parameters (Olden et al. 2008).

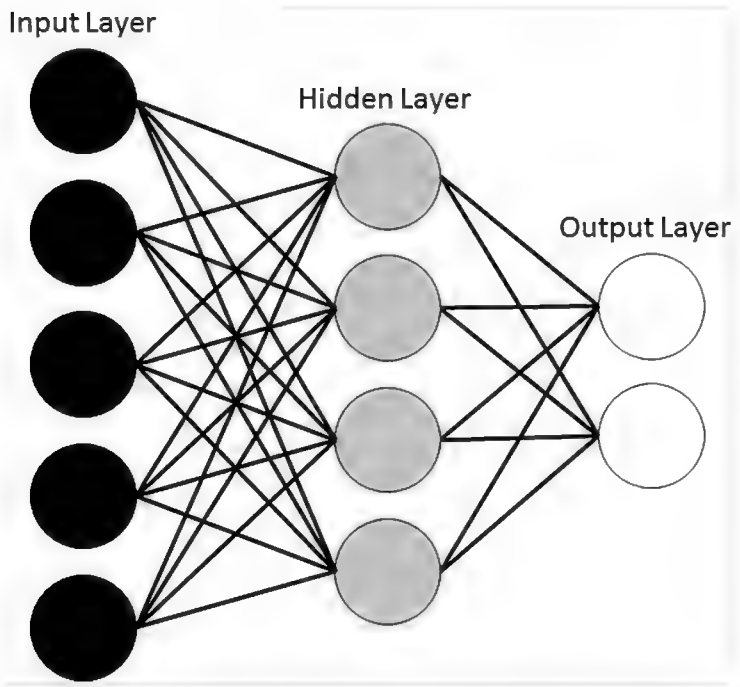


Figure 5.

**Artificial Neural Network Schematic**: A neural network is made up of three layers (input, hidden, output). Each layer contains interconnected units (neurons). Each connection has an assigned connection weight. The number of hidden units and the connection weights are iteratively improved to minimize the error between the output and the target.

ANN can be a powerful modeling tool when the underlying relationships are unknown and the data are imprecise and noisy (Lek and Guégan 1999). Interpretation of the ANN can be difficult and neural networks are often referred to as a "black box" method (Lek and Guégan 1999, Olden et al. 2008, Wieland and Mirschel 2008, Kampichler et al. 2010). ANNs can be more complicated to implement and are more computationally expensive than tree-based ML methods (Olden et al. 2008), but ANNs can accommodate a major gain in computational speed with a minor sacrifice in accuracy. For example, an ANN with one fourth the computational cost of a traditional satellite data retrieval algorithm (that uses an iterative method) has an accuracy nearly identical ($\pm$ 0.1) to the traditional algorithm (Young 2009). Overfitting can be a problem (Kampichler et al. 2010). Many ANNs mimic standard statistical methods (A. Fielding pers. comm.), so a good practice while using ANNs is to also include a rigorous suite of validation tests and a general linear model for comparison (Özesmi et al. 2006).

For a more detailed description of a multi-layer, feed-forward ANN with back propagation see section 4.1 in Kotsiantis et al. 2006 and section 5 in Bishop 2006. For more information on ANNs in general see Hagan et al. 2014.

## Support Vector Machines

A Support Vector Machine (SVM) is a type of binary classifier. Data are represented as points in space and classes are divided by a straight line "margin". The SVM maximizes the margin by placing the largest possible distance between the margin and the instances on both sides (Fig. 6a; Moguerza and Muñoz 2006, Rasmussen and Williams 2006, Zhao et al. 2008, Hsieh 2009, Kampichler et al. 2010, Zhao et al. 2011). A new data point would be classified according to which side of the margin it fell. SVMs are trained iteratively using the Sequential Minimal Optimization algorithm, which breaks the binary classification problem into several sub-problems and finds the maximum distance between the margin and the instances in each class (Keerthi and Gilbert 2002).

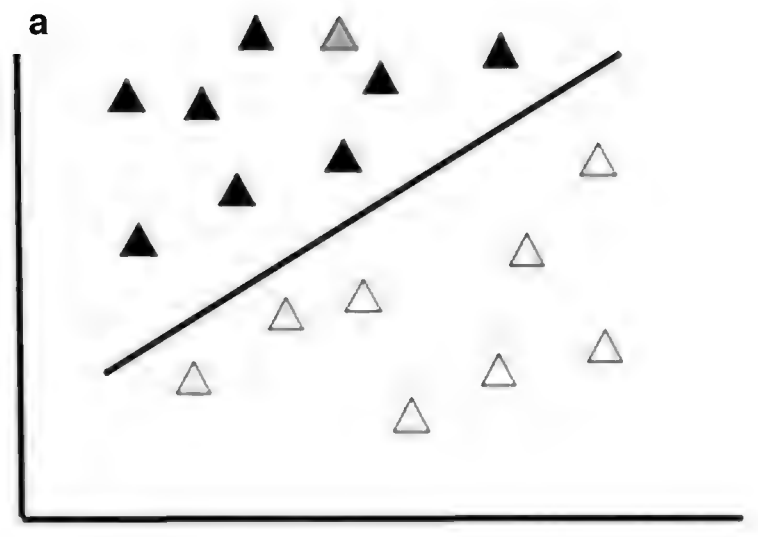


Figure 6.

**Support Vector Machine Schematic**

**a**: This simplified schematic shows the margin (black line) dividing the data set into two classes. A new datum (grey), will be classified according to its position relative to the margin.
**b**: If the data are noisy, and not easily separated, a “soft margin” (dotted lines) can be used to separate the two classes.

SVM is well suited for problems with many features compared to instances (Curse of Dimensionality) and is capable of avoiding problems with local minima (Kotsiantis et al.

2006). Because most real-world data cannot be separated with a straight line, significant additional processing may be required. If the classes overlap only slightly, a “buffer zone” or "soft margin" can be created around the hard decision boundary ( Fig. 6bVeropoulos et al. 1999). Another solution is to map the data onto a higher-dimensional feature space wherein a linear boundary can be found. An algorithm called a kernel function is used to translate data into the new feature space. Choosing the correct kernel function is important (Kotsiantis et al. 2006) and can slow the training process. SVMs are excellent binary classifiers when given labeled training data. Problems with more than one class, must be divided into multiple binary classification problems. When data are unlabeled, SVMs can be used for clustering, and this is called Support Vector Clustering (Ben-Hur et al. 2001).

For a more detailed description of SVMs see section 6 in Kotsiantis et al. 2006 and chapter 9 in James et al. 2013. For a discussion of kernel functions see Genton 2001.

**Genetic Algorithm**

Genetic Algorithms (GA) are based on the process of evolution in natural systems in that a population of competing solutions evolves over time to converge on an optimal solution (Holland 1975, Goldberg and Holland 1988, Koza 1992, Haupt and Haupt 2004, Olden et al. 2008). Solutions are represented as “chromosomes” and model parameters are represented as “genes” on those chromosomes (Fig. 7). Training a GA has four steps: 1) random potential solutions are generated (chromosomes), 2) potential solutions are altered using “mutation”, and “recombination”, 3) solutions are evaluated to determine fitness (minimizing error), and 4) the best solutions cycle back to step 2 (Holland 1975, Mitchell 1998, Haefner 2005). Each cycle represents a “generation”. Each chromosome is evaluated using a fitness function that scores its accuracy (Reeves and Rowe 2002). Depending on the nature of the problem the GA is trying to solve, the chromosome can be strings of bits, real values, rules, or permutations of elements (Recknagel 2001).

An advantage of GA is the removal of the often arbitrary process of choosing a model to apply to the data (Jeffers 1999) and can be used to find the characteristics of other ML techniques, such as the weights and architectures of ANNs (Siddique and Tokhi 2001, Yen and Lu 2000). GAs have seen a rise in popularity due to development of the Genetic Algorithm for Rule-Set Prediction (GARP) used to predict species distributions (Stockwell and Noble 1992). GAs are very popular in hydrology (see Mulligan and Brown 1998 for description of how GA was used to find the Pareto Front) and meteorology (Haupt 2009). GAs are able to cope with uneven sampling and small sample sizes (Olden et al. 2008). GAs were developed with broad application in mind and can use a wide range of model structures and model-fitting approaches (Olden et al. 2008). As a result, a larger burden is placed on the user to select complicated model parameters with little guidance, and the fixed-length “chromosomes” can limit the potential range of solutions (Olden et al. 2008). GAs are not best for all problems and many traditional statistical techniques can perform just as well or better (Olden et al. 2008). GARP, in particular, can be susceptible to overfitting (Lawler et al. 2006, Elith et al. 2008).

For a more detailed discussion of GA see Mitchell 1998 and Sastry et al. 2013.

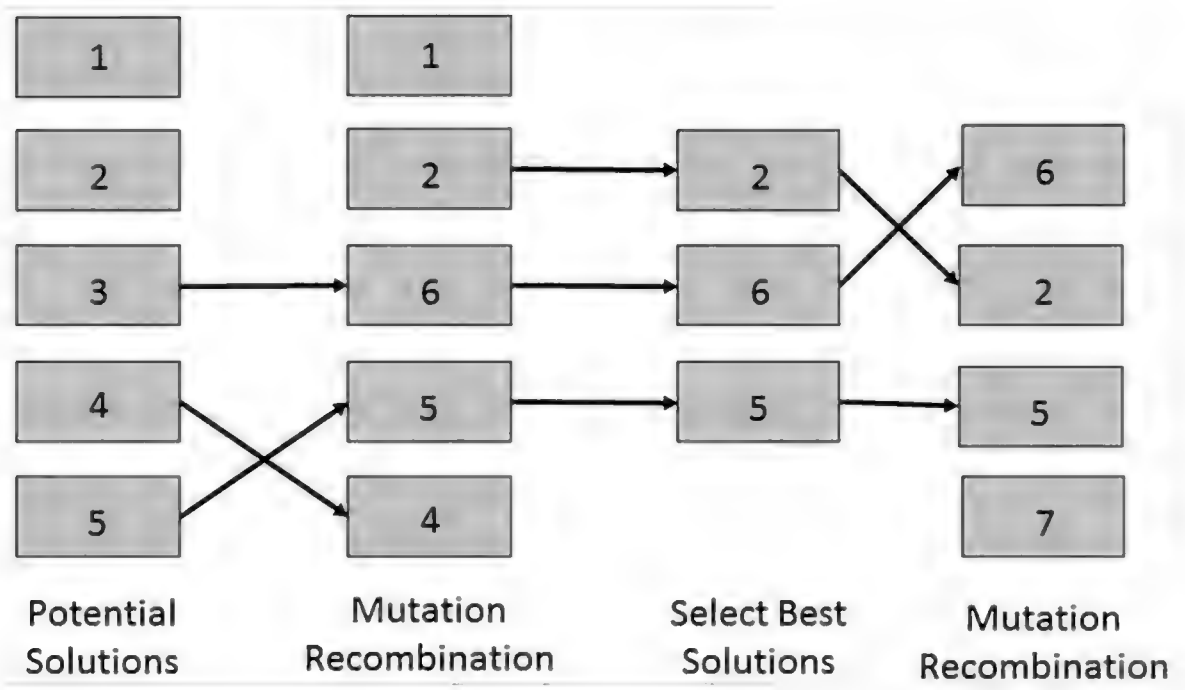


Figure 7.

**Genetic Algorithm Schematic**: In this simplified schematic of a genetic algorithm, the five potential solutions, or "chromosomes", undergo mutation and recombination. Then the best performing solutions are selected for another iteration of mutation and recombination. This cycle is repeated until an optimal solution is found.

## Fuzzy Inference Systems

Fuzzy inference methods, such as Fuzzy Logic and Inductive Logic Programming, provide a practical approach to automating complex analysis and inference in a long workflow (Williams et al. 2009). Given a set of training examples, a fuzzy inference system will find a set of rules that can be used for prediction of new instances. The output is in terms of a natural language set of "if/then" rules (Wieland 2008). For example, a set of rules that predict when a child receives an allowance might be "if the room is clean and the behavior is polite, then the allowance is dispensed" (Fig. 8). The if/then rules are created through an algorithm that iteratively selects each class and refines the if-statement until only the selected class remains (e.g. Džeroski 2009). Two examples of these algorithms are FOIL and PROGOL (Muggleton 1995, Quinlan and Cameron-Jones 1995. The National Center for Atmospheric Research (NCAR) has developed three fuzzy inference algorithms to address a complex problem in meteorology (Williams et al. 2009) and these methods have been used to predict landslide susceptibility (Pradhan 2010).

Fuzzy inference systems perform a rule induction task. The resulting rules can be easy to understand and interpret, as long as the rule sets are not too large, but overfitting can be a problem (Kampichler et al. 2010). Because fuzzy inference systems use a larger pool of possible rules and are more expressive, they can be more computationally demanding.

For more information about fuzzy inference systems and rule induction see Fürnkranz et al. 2012b and Wang et al. 2007.

| | | | |
|---|---|---|---|
| 1 | Clean | Polite | Dispensed |
| 2 | Dirty | Polite | Withheld |
| 3 | Dirty | Impolite | Withheld |
| 4 | Clean | Impolite | Withheld |
| 5 | Clean | Polite | Dispensed |
| 6 | Clean | Polite | Dispensed |

IF room = clean
AND child = polite
THEN allowance = dispensed
IF room = dirty
OR child = impolite
THEN allowance = withheld

Figure 8.

**Fuzzy Inference Rule Set**. This is a simplified example of an "if/then" rule set derived from data (in the table) using fuzzy inference. The inferred rules can be used to predict when an allowanced with be dispensed.

## Bayesian Methods

Bayesian ML methods are based on Bayesian statistical inference, which started in the 18th century with the development of Bayes' theorem (Laplace 1986). These methods are based on expressing the true state of the world in terms of probabilities and then updating the probabilities as evidence is acquired (Bishop 2006). In most cases, it is important to know the probability that a new datum belongs to a given class, not just the inferred class. The Bayesian approach can contribute to several ML and traditional statistical techniques, but this section will focus on Bayesian Classifiers. A Bayesian classifier calculates a probability density for each class (Fig. 9). The probability density is a curve showing, for any given value of the independent variable, the likelihood of being a member of that class (Fig. 9). The new datum is assigned to the class with the highest probability. The values of the independent variable that have an equal probability of being in either class are known as the decision boundary and this marks the dividing line between the classes. In the real world, it can be difficult to calculate these *a priori* probabilities and the user must often make a best-guess.

A Bayesian classifier gives good results in most cases and requires fewer training data compared to other ML methods (Kotsiantis et al. 2006). It is useful when there are more than two distinct classes. The disadvantage is that it can be very hard to specify prior probabilities and results can be quite sensitive to the selected prior. This method does assume that variables are independent, which is not always true (e.g., Lorena et al. 2011). Some Baysian classifiers have Gaussian assumptions which may not be reasonable for the problem at hand. Another issue is that if a specific feature never appears in a class, the resulting zero probability will complicate calculations; therefore, a small probability must often be added, even if the feature does not appear in the class.

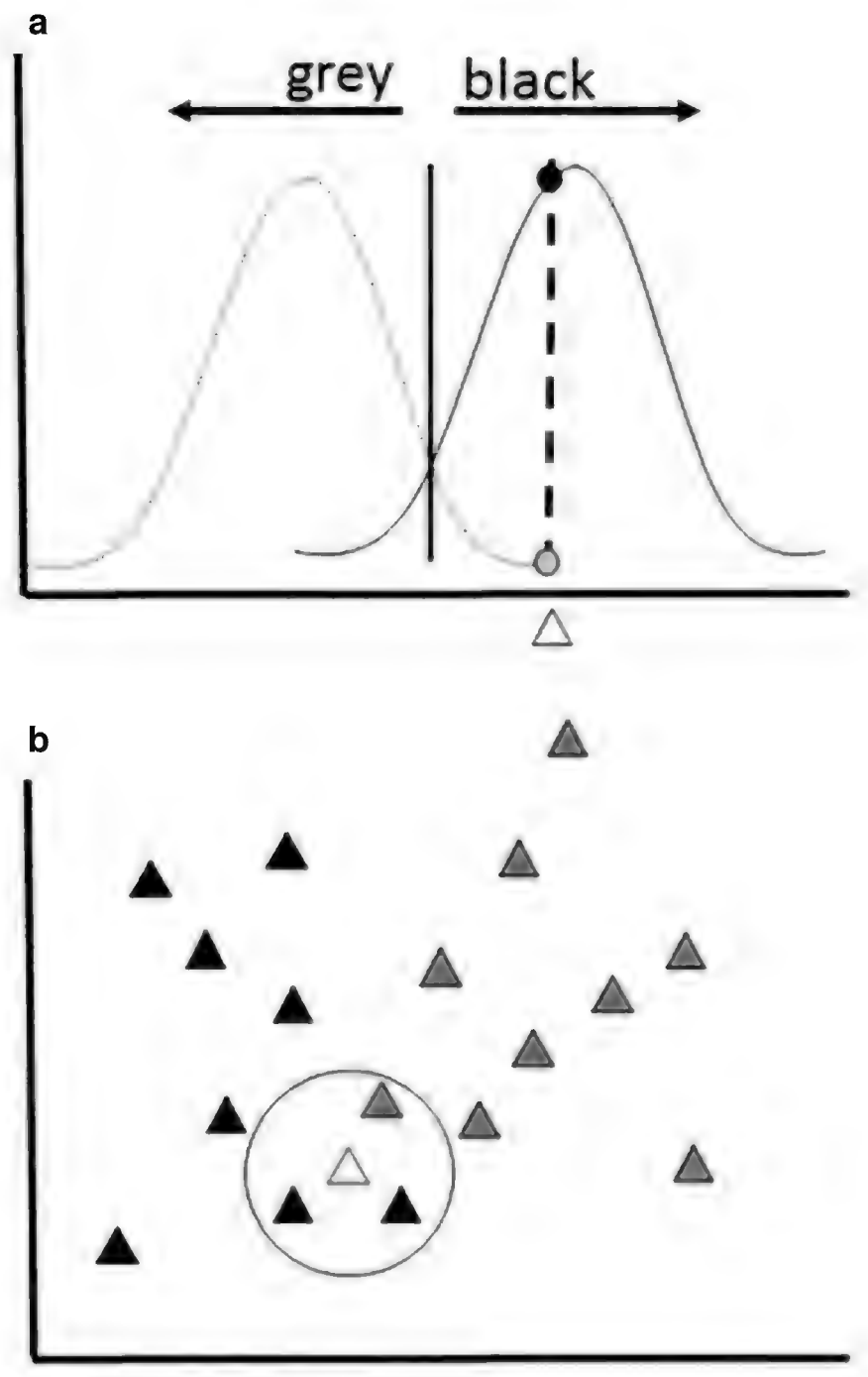


Figure 9.

**Bayesian Classifier Schematic**: This diagram shows a simplified schematic of a Bayesian classifier working to assign a new datum (white triangle) to one of two classes (grey and black).

**a**: Probability Density Plot: A Bayesian classifier calculates a probability density for each class (solid and dotted curve) across a range of values for the new datum (white triangle), which is classified according to which probability is highest at its value (black). The value for which the datum has an equal probability of being in both classes is called the decision boundary (black line).

**b**: Data Plot: An object to be classified (white) can belong to one of two groups (grey or black). This method would classify the object within the group with the highest probability of being correct. In this example, the white item would be classified as a member of the black group because the probability is higher (Black = 8/17 * 2/8 and Grey = 9/17 *1/9)

For a more detailed discussion of Bayesian Classifiers and Bayesian Networks see section 5.1 in Kotsiantis et al. 2006.

## Using ML in Ecology and Earth Science

For many researchers, machine learning is a relatively new paradigm that has only recently become accessible with the development of modern computing. While the adoption of ML methods in earth science and ecology has been slow, there are several published studies using ML in these disciplines (e.g., Park and Chon 2007). The following is a brief review of the different published applications of ML in earth science and ecology.

### Habitat Modeling and Species Distribution

Understanding the habitat requirements of a species is important for understanding its ecology and managing its conservation. Habitat modelers are interested in using multiple data sets to make predictions and classifications about habitat characteristics and where taxa are likely to be located or engaging in a specific behavior (e.g., nesting Fielding 1999b, Cutler et al. 2007). The rule-sets developed are referred to as Species Distribution Models (SDM) and can use a wide variety of ML methods to make their predictions or none at all (Guisan and Thuiller 2005). Typically, an algorithm would be trained using a data set matching environmental variables to taxon abundance or presence/absence data. If the algorithm tests well, it can be given a suite of environmental variables from a different location to make predictions about what taxa are present. This technique has been used to identify current suitable habitat for specific taxa, model future species distributions including predicting invasive and rare species presence, and predict biodiversity of an area (Tan and Smeins 1996, Kampichler et al. 2000, Cutler et al. 2007, Olden et al. 2008, Knudby et al. 2010a). Common tools include Random Forest (Cutler et al. 2007, Peters et al. 2007), classification and decision trees (Ribic and Ainley 1997, Bell 1999, Kobler and Adamic 2000, Vayssières et al. 2000, Debeljak et al. 2001, Miller and Franklin 2002), neural networks (MASTRORILLO et al. 1997, Guégan et al. 1998, Fielding 1999a, Manel et al. 1999, Brosse et al. 2001, Thuiller 2003, Dedecker et al. 2004, Segurado and Araújo 2004, Özesmi et al. 2006), genetic algorithms (D'Angelo et al. 1995, Stockwell and Peters 1999, Stockwell 1999, McKay 2001, Peterson et al. 2002, Wiley et al. 2003, Termansen et al. 2006), support vector machines (Pouteau et al. 2012), and Bayesian classifiers (Fischer 1990, Brzeziecki et al. 1993, Guisan and Zimmermann 2000).

### Species Identification

Identifying taxa can require specialized knowledge only possessed by a very few and the data set requiring expert curation can be large (e.g., automated collection of images and sounds). Thus, the expert annotation step is a major bottleneck in biodiversity studies. In order to increase throughput, algorithms are trained on images, sounds, and other types of data labeled with taxon names. (For more information about automated taxon identification specifically, see Edwards et al. 1987 and MacLeod 2007). The trained algorithms can then automatically annotate new data. This technique has been used to identify plankton, spiders, and shellfish larvae from images (Boddy and Morris 1999, Do et al. 1999, Sosik and Olson 2007, Goodwin et al. 2014). Bacterial taxa have been identified from gene sequences (Wang et al. 2007). Audio files of amphibian, bird, bat, insect, elephant, cetacean, and deer sounds have been classified to species (Parsons and Jones 2000,

Jennings et al. 2008, Chesmore 2004, Acevedo et al. 2009, Armitage and Ober 2010, Kasten et al. 2010). Fish and algal species have been identified using acoustic (Simmonds et al. 1996) and optical characteristics (Balfoort et al. 1992, Boddy et al. 1994). ML has been used to differentiate between the radar signals of birds and abiotic objects (Rosa et al. 2015). In some cases, individuals of the same species can be distinguished even if the individuals themselves are unknown *a priori* (Reby et al. 1998, Fielding 1999b). Common tools include support vector machines (Fagerlund 2007, Sosik and Olson 2007, Acevedo et al. 2009, Armitage and Ober 2010, Goodwin et al. 2014, Rosa et al. 2015), Random Forest (Armitage and Ober 2010, Rosa et al. 2015), Bayesian classifiers (Fielding 1999a, Wang et al. 2007), genetic algorithms (Jeffers 1999), and neural networks (Balfoort et al. 1992, Boddy et al. 1994, Simmonds et al. 1996, Do et al. 1999, Parsons and Jones 2000, Jennings et al. 2008, Armitage and Ober 2010, Rosa et al. 2015).

**Remote Sensing**

Satellite images and other data gathered from sensors at great elevation (e.g., LIDAR) are an excellent way to gather large amounts of data about Earth over broad spatial scales. In order to be useful, these data must go through some minimum level of processing (Atkinson and Tatnall 1997) and are often classified into land cover or land use categories (Guisan and Zimmermann 2000). ML methods have been developed to automate these laborious processes (Lees and Ritman 1991, Fitzgerald and Lees 1992, Lees 1996, Atkinson and Tatnall 1997, Guisan and Zimmermann 2000, Ham et al. 2005, Pal 2005, Gislason et al. 2006, Lakshmanan 2009). ML methods can be used to infer geophysical parameters from remote sensing data, such as inferring the Leaf Area Index from Moderate Resolution Imaging Spectrometer data (Rumelhart et al. 1986, Hsieh 2009, Krasnopolsky 2009). Sometimes remote sensing data and the parameters inferred from them can require spatial interpolation in the vertical or horizontal dimension, which is often performed using ML methods (Krasnopolsky 2009, Li et al. 2011). Common tools for classifying remote sensing images include Random Forest (Knudby et al. 2010b, Duro et al. 2012), support vector machines (Durbha et al. 2007, Knudby et al. 2010b, Zhao et al. 2011, Duro et al. 2012, Mountrakis et al. 2011), neural networks (Rogan et al. 2008), genetic algorithms (Haupt 2009), and decision trees (Huang and Jensen 1997). Random forest and support vector machines have been used for spatial interpolation of environmental variables (Li et al. 2011). Artificial neural networks have been used to infer geophysical parameters from remote sensing data (Hsieh 2009, Rumelhart et al. 1986, Krasnopolsky 2009).

**Resource Management**

Making decisions about conservation and resource management can be very difficult because there is often not enough data for certainty and the consequences of being wrong can be disastrous. ML methods can provide a means of increasing certainty and improving results, especially techniques that incorporate Bayesian probabilities. Several algorithms have been applied to water (Maier and Dandy 2000, Haupt 2009), soil (Henderson et al. 2005, Tscherko et al. 2007), and biodiversity/wildlife management (Baran et al. 1996, Lek et al. 1996b, Lek et al. 1996a, Giske et al. 1998, Guégan et al. 1998, Spitz and Lek 1999, Chen et al. 2000, Vander Zanden et al. 2004, Jones et al. 2006, Sarkar et al. 2006, Worner

and Gevrey 2006, Cutler et al. 2007, Quintero et al. 2014, Bland et al. 2014). ML methods have been used to model population dynamics, production, and biomass in terrestrial, aquatic, marine, and agricultural systems (Scardi 1996, Recknagel 1997, Scardi and Harding 1999, Recknagel et al. 2000, Schultz et al. 2000, Džeroski 2001, Recknagel et al. 2002, McKenna 2005, Muttil and Lee 2005). Some specific examples of ML applications in resource management and conservation include 1) inference of IUCN (International Union for Conservation of Nature) conservation status of Data Deficient species (Bland et al. 2014, Quintero et al. 2014), 2) predicting farmer risk preferences (Kastens and Featherstone 1996), 3) predicting the production and biomass of various animal populations (Brey et al. 1996), 4) examining the effect of urbanization on bird breeding (Lee et al. 2007), 5) predicting disease risk (Furlanello et al. 2003, Guo et al. 2005), and 6) modeling ecological niches (Drake et al. 2006). Being able to make these types of predictions and inferences can help focus conservation efforts for maximum impact (Knudby et al. 2010a, Guisan et al. 2013). Common ML methods for resource management include genetic algorithms (Haupt 2009), neural networks (Brey et al. 1996, Kastens and Featherstone 1996, Recknagel 1997, Giske et al. 1998, Guégan et al. 1998, Schultz et al. 2000, Lee et al. 2007), support vector machines (Guo et al. 2005, Drake et al. 2006), fuzzy inference systems (Tscherko et al. 2007), decision trees (Henderson et al. 2005, Jones et al. 2006), and Random Forest (Furlanello et al. 2003, Cutler et al. 2007, Quintero et al. 2014).

**Forecasting**

Discovery of deterministic chaos in meteorological models (Lorenz 1963) led to reconsideration of the use of traditional statistical methods in forecasting (Pasini 2009). Today, predictions about weather are often made using ML methods. The most common ML methods used in meteorological forecasting are genetic algorithms, which have been used to model rainy vs non-rainy days (Haupt 2009) and severe weather (Hsieh 2009). Forecasting can be important for applications other than weather prediction. In atmospheric science, neural networks are able to find dynamics hidden in noise and successfully forecast important variables in the atmospheric boundary layer (Pasini 2009). The oceanography community makes extensive use of neural networks for forecasting sea level, waves, and sea surface temperature (Wu et al. 2006, Hsieh 2009). In addition to being directly used for forecasting, neural networks are commonly used for downscaling environmental and model output data sets used in making forecasts (Casaioli et al. 2003, Hsieh and Hsieh 2003, Marzban 2003).

**Environmental Protection and Safety**

Just as ML can help resource managers make important decisions with or without adequate data coverage, environmental protection and safety decisions can be aided with ML methods when data are sparse. ML has been used to classify environmental samples into inferred quality classes in situations where direct analyses are too costly (Džeroski 2001). The mutagenicity, carcinogenicity, and biodegradability of chemicals have been predicted based on structure without lengthy lab work (Džeroski 2001). Sources of air contaminants have been identified and characterized in spite of lack of *a priori* knowledge

about source location, emission rate, and time of release (Haupt et al. 2009a). ML can relate pollution exposure to human health outcomes (Džeroski 2001). Common ML methods for environmental protection include genetic algorithms (Haupt et al. 2009a), Bayesian classifiers (Walley et al. 1992), neural networks (Ruck et al. 1993, Walley and S. 1996, Walley et al. 2000), and fuzzy inference systems (Srinivasan et al. 1997, Džeroski et al. 1999, Džeroski 2001).

### Climate Change Studies

One of the more pressing societal problems is the mitigation of and adaptation to climate change. Policy-makers require well-formed predictions in order to make decisions, but the complexity of the climate system, the interdisciplinary nature of the problem, and the data structures prevents the effective use of linear modeling techniques. ML is used to study important processes such as El Niño, the Quasi Biennial Oscillation, the Madden-Julian Oscillation, and monsoon modes (Cavazos et al. 2002, Hsieh 2009, Krasnopolsky 2009, Pasini 2009), and to predict climate change itself (Casaioli et al. 2003, Hsieh and Hsieh 2003, Marzban 2003, Pasini 2009). Predictions about the greenhouse effect (Seginer et al. 1994) and environmental change (Guisan and Zimmermann 2000) have also been made using ML. A very common use of ML in climate science is downscaling and post processing data from General Circulation Models (refs in Hsieh 2009, Pasini 2009). Ecological niche modeling and predictive vegetation mapping (as discussed above) can help predict adaptation to climate change (Wiley et al. 2003, Iverson et al. 2004). The most commonly used ML method in climate change studies is the neural network (Guisan and Zimmermann 2000, Pasini 2009).

## Discussion

### How can ML advance ecology and earth science?

The application of ML methods in ecology and earth science has already demonstrated the potential for increasing the quality and accelerating the pace of science. One of the more obvious ways ML does this is by coping with data gaps. The Earth is under-sampled, despite spending hundreds of millions of dollars on earth and environmental science (e.g., Webb et al. 2010). Where possible, ML allows a researcher to use data that are plentiful or easy to collect to infer data that are scarce or hard to collect (e.g., Wiley et al. 2003, Edwards et al. 2005, Buddemeier et al. 2008). Conservation managers are particularly well positioned to take advantage of ML via SDMs in invasive species management, critical habitat identification, and reserve selection (Guisan et al. 2013). Depending on the ML method used, one can also learn more about how a system works, for example through the Random Forest Variable Importance analysis. ML methods let the data tell the story and work backwards to understand the system while many numerical models impose a set of equations that may or may not be adequate. Another important way ML can fill in data gaps is through downscaling and performing spatial interpolation (Li et al. 2011). There will never be enough research funding to sample everything all of the time. ML can be a tractable method for addressing the data gaps that prevent scientific progress.

ML can accelerate the pace of science by quickly performing complex classification tasks normally performed by a human. A bottleneck in many ecology and earth science workflows are the manual steps performed by an expert, usually a classification task such as identifying a species. Expert annotation can be even more time consuming when the expert must search through a large volume of data, like a sensor stream, for a desired signal (Kasten et al. 2010). Rather than having all of the data classified by an expert, the expert only needs to review enough data to train and test an algorithm. This bottleneck has been addressed for some types of taxon identification (Cornuet et al. 1996, Sosik and Olson 2007, Acevedo et al. 2009, Armitage and Ober 2010), finding relevant data in sensor streams (Kasten et al. 2010), and building a reference knowledgebase for image analysis (Huang and Jensen 1997). In addition to relieving a bottleneck, ML methods can sometimes perform tasks more consistently than experts, especially when there are many categories and the task continues over a long period of time (Culverhouse et al. 2003, Jennings et al. 2008). In these cases, ML methods can improve the quality of science by providing more quantitative and consistent data (Sutherland et al. 2004, Olden et al. 2008, Acevedo et al. 2009).

As discussed above, ML techniques can perform better than traditional statistical methods in some systems, but a direct comparison of performance between ML techniques and traditional statistical methods is difficult because there is no universal measure of performance and results can be very situation-specific (Fielding 2007). The true measure of the utility of a tool is how well it can make predictions from new data and how well it can be generalized to new situations. Highly significant p-values, R2 values, and accuracy measurements may not reflect this. A study comparing 33 classification methods (including ML and traditional statistics) with 32 data sets found no real difference in performance and suggested that choice of algorithm be driven by factors other than accuracy, such as the characteristics of the data set (Lim et al. 2000). If the accuracy is not significantly improved using ML, it may be better to use a traditional method that is more familiar and accepted by peers and managers. Best practice is to test multiple methods (including traditional statistics) while probing the trade-off between bias and accuracy and choose the technique that best fits the problem. In many natural systems, where non-linear and interaction effects are common, a ML-based model may be more useful. Individual researchers need to select a method based on the specific problem and the data at hand.

## Why don't more people use ML?

Even though ML can outperform traditional statistics in some applications (Manel et al. 1999, Kampichler et al. 2000, Segurado and Araújo 2004, Elith et al. 2006, Peters et al. 2007, Pasini 2009, Armitage and Ober 2010, Knudby et al. 2010a, Li et al. 2011, Zhao et al. 2011, Bhattacharya 2013), the potential of ML methods in ecology and earth science has not been exhausted (Olden et al. 2008). The reasons for this are social and technical. New methods can be resisted by established scientists, which can delay wide-spread use (Azoulay et al. 2015). ML methods (as well as some more complex statistical models) can require a high degree of math skill to understand in detail, which means either a long familiarization phase or an acceptance of the algorithm as a "black box" (Kampichler et al.

2010). ML methods are highly configurable; thus, it can be overwhelming for researchers to choose the proper test for the job (Kampichler et al. 2010). Many of them require programming skills (e.g. scikit-learn) that many ecologists lack (Olden et al. 2008); however, tools like MatLab and R have developed more user-friendly interfaces and lowered the barrier to adoption for many users. Alternatively, many of the traditional statistical methods are fast to calculate and give easy-to-interpret metrics, like p-values (Olden et al. 2008, Kampichler et al. 2010). Traditional statistical methods are easier to find as a part of an off-the-shelf software package with a user interface and much of the complicated inner workings pleasantly hidden. Traditional statistical methods are part of a typical graduate and undergraduate education in the sciences whereas ML techniques are not. All of these make ML methods less attractive to practicing natural scientists than traditional statistical methods.

Another barrier to using ML techniques is the need for adequate amounts of training and test data within the desired range of prediction. This places an important constraint on the application of ML to problems that have appropriate, annotated data sets available. For example, the Google image recognition algorithm was developed using 1.2 million annotated images (Simonite 2016). Rarely does a natural science domain have a quality, annotated data set that large. In addition, the validity of a ML model is restricted to the range represented by the training and test data. For example, a bird behavior model that was trained only on data collected during the summer will not be able to predict winter behavior. This is an important problem in the natural sciences, where the need for extrapolation is high (e.g. predicting climate change). The lack of high-quality data for model development has been cited as a major bottleneck in many fields of ML application (e.g. Bewley et al. 2015, Thessen and Patterson 2011). There are techniques available for developing a model when the data set is small (Corkill and Gormley 2016), but some methods (e.g., cross-validation, Bayesian) give a weaker estimate of model error (Guisan and Zimmermann 2000, Hsieh 2009). Traditional statistical methods are validated using p values and tend to require much less data to develop a useful model. Thus, it can be harder to validate a ML model than a traditional statistical model.

This combination of lack of formal education and important data restrictions can lead to naive applications of ML techniques, which can increase resistence to their adoption.

Finally, communication between the ML research community and natural science research community is poor (Wagstaff 2012). The financial sector is applying ML, suggesting that communication is possible when the potential monetary reward is great enough. (The application of ML to the financial sector has had mixed results and uses only some of the same ML techniques discussed herein Fletcher 2016.) There is too much reliance on abstract metrics in the ML research community and not enough consideration of whether or not a particular ML advance will result in a real-world impact (Wagstaff 2012). The small community of ecologists using ML to develop SDMs are not communicating the value of their research to decision-makers and accounts of SDMs being used successfully in conservation are hidden in grey literature (Guisan et al. 2013). Communication and collaboration between the ML community, the ecology community, and the earth science community is poor.

## Next Steps

How can the use of ML methods in ecology and earth science be encouraged? One barrier that has been partially lowered is the lack of tools and services to support the application of ML in these domains. Use of ML algorithms built with user infrastructure, such as GRASS-GIS (Garzón et al. 2006) and GARP (Stockwell and Noble 1992) is higher than algorithms without such infrastructure. ML capabilities in R and MatLab have continued to make these methods more user-friendly. Programming skills have become more common in the natural sciences, but user interfaces are still very important for adoption of techniques.

Research scientists want to have a good understanding of the algorithms they use, which makes adoption of a new method a non-trivial investment. Reducing the cost of this investment for ML techniques is an important part of encouraging adoption. One way to do this is through a trusted collaborator who can simultaneously guide the research and transfer skills. These collaborators can be difficult to find, but many potential partners can be found in industry. A useful tool would be a publicly-available repository of annotated data sets to act as a sandbox for researchers wanting to learn and experiment with these methods, similar to Kaggle (https://www.kaggle.com/) but with natural science data. Random Forest is easier for a beginner to implement, gives easy to interpret results, and has high performance on ecology and earth science classification problems (Prasad et al. 2006, Kampichler et al. 2010); thus, Random Forest would be a good starting point for a ML novice. Students can be exposed to ML and command line programming through their graduate education, eliminating the need for a costly time investment during their research career. In addition, an improved statistical education for students would make them more aware of the limitations imposed by rigid models and thus more open to trying ML for some problems. An important part of promoting new techniques is recognizing the practical needs of researchers and working within those boundaries to facilitate change.

Finally, ML successes and impacts in ecology and earth science need to be more effectively communicated and the results from ML analyses need to be easily interpreted for decision-makers (Guisan et al. 2013). Research communities need to do a better job of communicating across domains about the impact of their results (Wagstaff 2012). For best communication between experts, collaborations should begin during and even before algorithm development to help properly define the problem being addressed, instead of developing an algorithm in isolation (Guisan et al. 2013). Once an algorithm has been successfully used in a decision-making process, the results need to be reported as a part of the published literature in addition to the grey literature.

Funding agencies can facilitate this process by specifically soliciting new collaborative projects (research projects, workshops, hack-a-thons, conference sessions) that apply ML methods to ecology and earth science in innovative ways and initiatives to develop education materials for natural science students. Proper implementation of ML methods requires an understanding of the data science and the discipline that can best be achieved through interdisciplinary collaboration.

## Conclusions

ML methods offer a diverse array of techniques, now accessible to individual researchers, that are well suited to the complex data sets coming from ecology and earth science. These methods have the potential to improve the quality of scientific research by providing more accurate models and accelerate progress in science by widening bottlenecks, filling data gaps, and enhancing understanding of how systems work. Application of these methods within the ecology and earth science domain needs to increase if society is to see the benefit. Adoption can be promoted through interdisciplinary collaboration, increased communication, increased formal and informal education, and financial support for ML research. Partnerships with companies interested in environmental issues can be an excellent source of knowledge transfer. A good introductory ML method is Random Forest, which is easy to implement and gives good results. However, ML methods have limitations and are not the answer to all problems. In some cases traditional statistical approaches are more appropriate (Meynard and Quinn 2007, Olden et al. 2008). ML methods should be used with discretion.

There are many more types of ML methods and subtly different techniques than what has been discussed in this paper. Implementing ML effectively requires additional background knowledge. A very helpful series of lectures by Stanford Professors Trevor Hastie and Rob Tibshirani called “An Introduction to Statistical Learning with Applications in R” can be accessed online for free and gives a general introduction to traditional statistics and some ML methods. Kaggle (https://www.kaggle.com/) is an excellent source of independent, hands-on data science lessons. A suggested introductory text is "Machine Learning Methods in the Environmental Sciences", by William Hsieh (Hsieh 2009), written at the graduate student level. A useful paper and book written for ecologists is "Machine learning methods without tears: A primer for ecologists" by Olden et al. (Olden et al. 2008) and “Machine Learning Methods for Ecological Applications” edited by Fielding (Fielding 1999a). ML can be mastered by natural scientists and the time invested in learning it can have significant reward.

## Acknowledgements

The author would like to acknowledge NASA for financial support and the Boston Machine Learning Meetup Group for inspiration. This paper was greatly improved by comments from Ronny Peters (reviewer), Christopher W. Lloyd, Holly A. Bowers, Alan H. Fielding, and Joseph Gormley.

## References


- Acevedo M, Corrada-Bravo C, Corrada-Bravo H, Villanueva-Rivera L, Aide TM (2009) Automated classification of bird and amphibian calls using machine learning: A comparison of methods. Ecological Informatics 4 (4): 206-214. DOI: 10.1016/j.ecoinf.2009.06.005
- Alpaydin E (2014) Introduction to Machine Learning. The MIT Press, 615 pp.
- Armitage D, Ober H (2010) A comparison of supervised learning techniques in the classification of bat echolocation calls. Ecological Informatics 5 (6): 465-473. DOI: 10.1016/j.ecoinf.2010.08.001
- Atkinson PM, Tatnall ARL (1997) Introduction Neural networks in remote sensing. International Journal of Remote Sensing 18 (4): 699-709. DOI: 10.1080/014311697218700
- Azoulay P, Fons-Rosen C, Graff Zivin J (2015) Does science advance one funeral at a time? National Bureau of Economic Research 21788: 1. DOI: 10.3386/w21788
- Balfoort HW, Snoek J, Smiths JR, Breedveld LW, Hofstraat JW, Ringelberg J (1992) Automatic identification of algae: neural network analysis of flow cytometric data. J Plankton Res 14 (4): 575-589. DOI: 10.1093/plankt/14.4.575
- Baran P, Lek S, Delacoste M, Belaud A (1996) Stochastic models that predict trouts population densities or biomass on macrohabitat scale. Hydrobiologia 337: 1-9.
- Bell J (1999) Tree-based methods. In: Fielding AH (Ed.) Machine Learning Methods for Ecological Applications. Springer US, New York, 89-106 pp.
- Ben-Hur A, Horn D, Siegelmann HT, Vapnik V (2001) Support vector clustering. Journal of Machine Learning Research 2: 125-137.
- Bewley M, Friedman A, Ferrari R, Hill N, Hovey R, Barrett N, Pizarro O, Figueira W, Meyer L, Babcock R, Bellchambers L, Byrne M, Williams S (2015) Australian sea-floor survey data, with images and expert annotations. Scientific Data 2: 150057. DOI: 10.1038/sdata.2015.57
- Bhattacharya M (2013) Machine learning for bioclimatic modelling. Int J Adv Comput Sci Appl 4 (2): 8.
- Bishop C (2006) Pattern Recognition and Machine Learning. Springer-Verlag, New York, 738 pp. [ISBN 978-0-387-31073-2]
- Bland L, Collen B, Orme CDL, Bielby J (2014) Predicting the conservation status of data-deficient species. Conservation Biology 29 (1): 250-259. DOI: 10.1111/cobi.12372
- Boddy L, Morris C (1999) Artificial neural networks for pattern recognition. In: Fielding AH (Ed.) Machine Learning Methods for Ecological Applications. Springer US, New York, 37-88 pp.
- Boddy L, Morris CW, Wilkins MF, Tarran GA, Burkill PH (1994) Neural network analysis of flow cytometric data for 40 marine phytoplankton species. Cytometry 15 (4): 283-293. DOI: 10.1002/cyto.990150403
- Breiman L (1996) Bagging predictors. Machine Learning 24 (2): 123-140. DOI: 10.1023/a:1018054314350
- Breiman L (2001a) Statistical modeling: The two cultures. Stat Sci 16 (3): 199-231.
- Breiman L (2001b) Random Forests. Machine Learning 45 (1): 5-32. DOI: 10.1023/a:1010933404324

- Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and Regression Trees. Chapman Hall/CRC Press, 368 pp. [ISBN 0412048418]
- Brey T, Jarre-Teichmann A, Borlich O (1996) Artificial neural network versus multiple linear regression: Predicting P/B ratios from empirical data. Marine Ecology Progress Series 140: 251-256.
- Brosse S, Lek S, Townsend C (2001) Abundance, diversity, and structure of freshwater invertebrates and fish communities: An artificial neural network approach. New Zealand Journal of Marine and Freshwater Research 35 (1): 135-145. DOI: 10.1080/00288330.2001.9516983
- Brzeziecki B, Kienast F, Wildi O (1993) A simulated map of the potential natural forest vegetation of Switzerland. J Veg Sci 4 (4): 499-508. URL: http://www.jstor.org/stable/3236077
- Buddemeier R, Smith S, Swaney D, Crossland C, Maxwell B (2008) Coastal typology: An integrative "neutral" technique for coastal zone characterization and analysis. Estuarine and Coastal Shelf Science 77 (2): 197-205. DOI: 10.1016/j.ecss.2007.09.021
- Camargo L, Yoneyama T (2001) Specification of training sets and the number of hidden neurons for multilayer perceptrons. Neural Comput 13 (12): 2673-2680.
- Casaioli M, Mantovani R, Proietti Scorzoni F, Puca S, Speranza A, Tirozzi B (2003) Linear and nonlinear postprocessing of numerical forecasted surface temperature. Nonlinear Process Geophys 10: 373-383.
- Cavazos T, Comrie A, Liverman D (2002) Intraseasonal variability associated with wet monsoons in southeast Arizona. Journal of Climate 2002 (15): 2477-2490. DOI: 10.1175/1520-0442(2002)0152.0.CO;2
- Chawla N (2005) Data mining for imbalanced datasets: An overview. In: Maimon O, Rokach L (Eds) The Data Mining and Knowledge Discovery Handbook. Springer, 853-867 pp. DOI: 10.1007/0-387-25465-X_40
- Chen DG, Hargreaves NB, Ware DM, Liu Y (2000) A fuzzy logic model with genetic algorithm for analyzing fish stock-recruitment relationships. Can. J. Fish. Aquat. Sci. 57 (9): 1878-1887. DOI: 10.1139/f00-141
- Chesmore D (2004) Automated bioacoustic identification of species. Ann Brazilian Acad Sci 76 (2): 435-440.
- Chon T, Park Y, Moon K, Cha E (1996) Patternizing commuinities by using an artificial neural network. Ecological Modelling 90: 69-78. DOI: 10.1016/0304-3800(95)00148-4
- Corkill DD, Gormley J (2016) Intelligent Strategies for Effective Quantitative Structure-Toxicity Relationship (QSTR) Screening. in proof 1: 1.
- Cornuet J, Aulagnier S, Lek S, Franck P, Solignac M (1996) Classifying individuals among infraspecific taxa using microsatellite data and neural networks. Comptes Rendus l'Académie des Sci Série III, Sci la vie 319 (12): 1167-1177.
- Culverhouse P, Williams R, Reguera B, Herry V, González-Gil S (2003) Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. Marine Ecology Progress Series 247: 17-25. DOI: 10.3354/meps247017
- Cutler DR, Edwards T, Beard K, Cutler A, Hess K, Gibson J, Lawler J (2007) Random forests for classification in ecology. Ecology 88 (11): 2783-2792. DOI: 10.1890/07-0539.1
- D'Angelo D, Meyer J, Howard L, Gregory S, Ashkenas L (1995) Ecological uses for genetic algorithms: predicting fish distributions in complex physical habitats. Can. J. Fish. Aquat. Sci. 52 (9): 1893-1908. DOI: 10.1139/f95-782

- De'ath G (2007) Boosted trees for ecological modeling and prediction. Ecology 81: 243-251. DOI: 10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2
- De'ath G, Fabricius K (2000) Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology 81 (11): 3178-3192.
- Debeljak M, Džeroski S, Jerina K, Kobler A, Adamič M (2001) Habitat suitability modelling for red deer (Cervus elaphus L.) in South-central Slovenia with classification trees. Ecol Modell 138: 321-330. DOI: 10.1016/S0304-3800(00)00411-7
- Dedecker AP, Goethals PM, Gabriels W, De Pauw N (2004) Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium). Ecological Modelling 174: 161-173. DOI: 10.1016/j.ecolmodel.2004.01.003
- Do M, Harp J, Norris K (1999) A test of a pattern recognition system for identification of spiders. Bull Entomol Res 89: 217-224.
- Domingos P (2012) A few useful things to know about machine learning. Communications of the ACM 55 (10): 78-87. DOI: 10.1145/2347736.2347755
- Drake J, Randin C, Guisan A (2006) Modelling ecological niches with support vector machines. J Appl Ecology 43 (3): 424-432. DOI: 10.1111/j.1365-2664.2006.01141.x
- Du K- (2010) Clustering: A neural network approach. Neural Networks 23 (1): 89-107. DOI: 10.1016/j.neunet.2009.08.007
- Durbha S, King R, Younan N (2007) Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. Remote Sensing of Environment 107: 348-361. DOI: 10.1016/j.rse.2006.09.031
- Duro D, Franklin S, Dubé M (2012) A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. Remote Sensing of Environment 118: 259-272. DOI: 10.1016/j.rse.2011.11.020
- Džeroski S (2001) Applications of symbolic machine learning to ecological modelling. Ecol Modell 146: 263-273. DOI: 10.1016/S0304-3800(01)00312-X
- Džeroski S (2009) Machine learning applications in habitat suitability modeling. In: Haupt S, Pasini A, Marzban C (Eds) Artificial Intelligence Methods in the Environmental Sciences. Springer Netherlands, Amsterdam, 397-412 pp.
- Džeroski S, Blockeel H, Kompare B, Kramer S, Pfahringer B, Van Laer W (1999) Experiments in predicting biodegradability. In: Džeroski S, Flach P (Eds) Proceedings of the Ninth International Conference on Inductive Logic Programming. Springer, 80-81 pp.
- Edwards M, Morse D, Fielding A (1987) Expert systems: frames, rules or logic for species identification? Bioinformatics 3 (1): 1-7. DOI: 10.1093/bioinformatics/3.1.1
- Edwards T, Cutler DR, Zimmermann N, Geiser L, Alegria J (2005) Model-based stratifications for enhancing the detection of rare ecological events. Ecology 86 (5): 1081-1090. DOI: 10.1890/04-0608
- Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. Journal of Animal Ecology 77 (4): 802-813. DOI: 10.1111/j.1365-2656.2008.01390.x
- Elith J, H. Graham C, P. Anderson R, Dudík M, Ferrier S, Guisan A, J. Hijmans R, Huettmann F, R. Leathwick J, Lehmann A, Li J, G. Lohmann L, A. Loiselle B, Manion G, Moritz C, Nakamura M, Nakazawa Y, McC. M. Overton J, Townsend Peterson A, J. Phillips S, Richardson K, Scachetti-Pereira R, E. Schapire R, Soberón J, Williams S, S. Wisz M, E. Zimmermann N (2006) Novel methods improve prediction of species'

distributions from occurrence data. Ecography 29 (2): 129-151. DOI: 10.1111/j.2006.0906-7590.04596.x

- Escalante HJ (2005) A Comparison of Outlier Detection Algorithms for Machine Learning. CiteSeerx 2005: e. DOI: 10.1.1.61.5991
- Fagerlund S (2007) Bird Species Recognition Using Support Vector Machines. EURASIP Journal on Advances in Signal Processing 2007 (1): 038637. DOI: 10.1155/2007/38637
- Fidelis M, Lopes H, Freitas A (2000) Discovering comprehensible classification rules using a genetic algorithm. Proceedings of CEC-2000, conference on evolutionary computation., 1. La Jolla, USA. 805-811 pp.
- Fielding AH (1999a) Machine Learning Methods for Ecological Applications. Springer US, New York, 261 pp.
- Fielding AH (1999b) An introduction to machine learning methods. In: Fielding AH (Ed.) Machine Learning Methods for Ecological Applications. Springer US, New York, 1-36 pp.
- Fielding AH (2007) Cluster and Classification Techniques in the BioSciences. Cambridge University Press, Cambridge, 260 pp. [ISBN 9780521618007]
- Fischer H (1990) Simulating the distribution of plant communities in an alpine landscape. Coenoses 5: 37-43.
- Fitzgerald R, Lees B (1992) The application of neural networks to the floristic classification of remote sensing and GIS data in complex terrain. In: American Society of Photogrametry and Remote Sensing Proceedings of the XVII Congress ASPRS. 570-573 pp.
- Fletcher T (2016) Machine Learning for Financial Market Prediction. University College London, Department of Computer Science, 207 pp.
- Forget G, Campin J-, Heimbach P, Hill CN, Ponte RM, Wunsch C (2015) ECCO version 4: an integrated framework for non-linear inverse modeling and global ocean state estimation. Geoscientific Model Development Discussions 8 (5): 3653-3743. DOI: 10.5194/gmdd-8-3653-2015
- Furlanello C, Neteler M, Merler S, Menegon S, Fontanari S, Donini A, Rizzoli A, Chemini C (2003) GIS and the random forest predictor: Integration in R for tick-borne disease risk assessment. In: Hornik K, Leisch F, Zeileis A (Eds) Proceedings of the 3rd International Workshop on Distributed Statistical Computing.
- Fürnkranz J, Gamberger D, Lavrač N (2012a) Rule Learning in a Nutshell. In: Fürnkranz J, Gamberger D, Lavrač N (Eds) Foundations of Rule Learning. Springer, 19-55 pp.
- Fürnkranz J, Gamberger D, Lavrač N (2012b) Foundations of Rule Learning. Springer, 334 pp. DOI: 10.1007/978-3-540-75197-7
- Gamberger D, Sekusak S, Medven Z, Sabljic A (1996) Application of Artificial Intelligence in Biodegradation Modelling. In: Peijnenburg WGM, Damborský J (Eds) Biodegradability Prediction. Springer, 41-50 pp. DOI: 10.1007/978-94-011-5686-8_5
- Gantayat SS, Misra A, Panda BS (2014) A Study of Incomplete Data – A Review. In: Satapathy SC, Udgata SK, Biswal BN (Eds) Advances in Intelligent Systems and Computing. Springer, 563 pp. URL: http://dx.doi.org/10.1007/978-3-319-02931-3_45 DOI: 10.1007/978-3-319-02931-3_45
- Garzón MB, Blazek R, Neteler M, Dios RSd, Ollero HS, Furlanello C (2006) Predicting habitat suitability with machine learning models: The potential area of Pinus sylvestris L.

in the Iberian Peninsula. Ecological Modelling 197: 383-393. DOI: 10.1016/j.ecolmodel.2006.03.015

- Genton M (2001) Classes of Kernels for Machine Learning: A Statistics Perspective. Journal of Machine Learning Research 2: 299-312.
- Giske J, Huse G, Fiksen O (1998) Modelling spatial dynamics of fish. Reviews in Fish Biology and Fisheries 8 (1): 57-91. DOI: 10.1023/a:1008864517488
- Gislason PO, Benediktsson JA, Sveinsson J (2006) Random Forests for land cover classification. Pattern Recognition Letters 27 (4): 294-300. DOI: 10.1016/j.patrec.2005.08.011
- Goldberg D, Holland J (1988) Genetic algorithms and machine learning. Machine Learning 3: 95-99. DOI: 10.1023/a:1022602019183
- Goodwin J, North E, Thompson C (2014) Evaluating and improving a semi-automated image analysis technique for identifying bivalve larvae. Limnol. Oceanogr. 12 (8): 548-562. DOI: 10.4319/lom.2014.12.548
- Guégan J, Lek S, Oberdorff T (1998) Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature 391 (6665): 382-384. DOI: 10.1038/34899
- Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecology Letters 8 (9): 993-1009. DOI: 10.1111/j.1461-0248.2005.00792.x
- Guisan A, Zimmermann N (2000) Predictive habitat distribution models in ecology. Ecol Modell 135 (2): 147-186. DOI: 10.1016/S0304-3800(00)00354-9
- Guisan A, Tingley R, Baumgartner J, Naujokaitis-Lewis I, Sutcliffe P, Tulloch AT, Regan T, Brotons L, McDonald-Madden E, Mantyka-Pringle C, Martin T, Rhodes J, Maggini R, Setterfield S, Elith J, Schwartz M, Wintle B, Broennimann O, Austin M, Ferrier S, Kearney M, Possingham H, Buckley Y (2013) Predicting species distributions for conservation decisions. Ecology Letters 16 (12): 1424-1435. DOI: 10.1111/ele.12189
- Guo Q, Kelly M, Graham C (2005) Support vector machines for predicting distribution of Sudden Oak Death in California. Ecological Modelling 182 (1): 75-90. DOI: 10.1016/j.ecolmodel.2004.07.012
- Haefner J (2005) Modeling Biological Systems: Principles and Applications. Springer US, New York, 475 pp. DOI: 10.1007/b106568
- Hagan MT, Demuth HB, Beale MH, Jesus Od (2014) Neural Network Design. Martin Hagan, 1012 pp.
- Ham J, Yangchi Chen, Crawford MM, Ghosh J (2005) Investigation of the random forest framework for classification of hyperspectral data. IEEE Trans. Geosci. Remote Sensing 43 (3): 492-501. DOI: 10.1109/tgrs.2004.842481
- Haupt R, Haupt S (2004) Practical Genetic Algorithms. John Wiley & Sons Inc., 253 pp. DOI: 10.1002/0471671746
- Haupt S (2009) Environmental optimization: Applications of genetic algorithms. In: Haupt S, Pasini A, Marzban C (Eds) Artificial Intelligence Methods in the Environmental Sciences. Springer Netherlands, Amsterdam, 379-396 pp.
- Haupt S, Allen C, Young G (2009a) Addressing air quality problems with genetic algorithms: A detailed analysis of source characterization. In: Haupt S, Pasini A, Marzban C (Eds) Artificial Intelligence Methods in the Environmental Sciences. Springer Netherlands, Amsterdam, 269-296 pp.

- Haupt S, Pasini A, Marzban C (Eds) (2009b) Artificial Intelligence Methods in the Environmental Sciences. Springer Netherlands, Amsterdam, 424 pp. DOI: 10.1007/978-1-4020-9119-3
- Hawkins D (2004) The Problem of Overfitting. Journal of Chemical Information and Computer Sciences 44 (1): 1-12. DOI: 10.1021/ci0342472
- Henderson BL, Bui EN, Moran CJ, Simon DA (2005) Australia-wide predictions of soil properties using decision trees. Geoderma 124: 383-398. DOI: 10.1016/j.geoderma.2004.06.007
- Holland J (1975) Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. MIT Press, Cambridge, 211 pp.
- Hsieh W (2009) Machine Learning Methods in the Environmental Sciences. Cambridge University Press, Cambridge, 349 pp. [ISBN 978-0-521-79192-2]
- Hsieh Y, Hsieh W (2003) An adaptive nonlinear MOS scheme for precipitation forecasts using neural networks. Weather Forecast 18: 303-310. DOI: 10.1175/1520-0434(2003)0182.0.CO;2
- Huang X, Jensen J (1997) A machine-learning approach to automated knowledge-base building for remote sensing image analysis with GIS data. Photogramm Eng Remote Sens 63 (10): 1185-1194.
- Iverson L, Prasad A, Liaw A (2004) New machine learning tools for predictive vegetation mapping after climate change: Bagging and random forest perform better than regression tree analysis. In: Smithers R (Ed.) Landscape Ecology of Trees and Forests, Proceedings of the Twelfth Annual IALE(UK) Conference. International Association for Landscape Ecology, 317-320 pp.
- Jain A (2010) Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31 (8): 651-666. DOI: 10.1016/j.patrec.2009.09.011
- James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning. Springer, 426 pp.
- Japkowicz N, Stephen S (2002) The class imbalance problem: A systematic study. Intelligent Data Analysis 6 (5): 429-449.
- Jeffers J (1999) Genetic Algorithms I. In: Fielding AH (Ed.) Machine Learning Methods for Ecological Applications. Springer Netherlands, Amsterdam, 107-122 pp.
- Jennings N, Parsons S, Pocock MJ (2008) Human vs. machine: identification of bat species from their echolocation calls by humans and by artificial neural networks. Can. J. Zool. 86 (5): 371-377. DOI: 10.1139/z08-009
- Jerez J, Molina I, García-Laencina P, Alba E, Ribelles N, Martín M, Franco L (2010) Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine 50 (2): 105-115. DOI: 10.1016/j.artmed.2010.05.002
- Jones M, Fielding A, Sullivan M (2006) Analysing Extinction Risk in Parrots using Decision Trees. Biodivers Conserv 15 (6): 1993-2007. DOI: 10.1007/s10531-005-4316-1
- Kampichler C, Džeroski S, Wieland R (2000) Application of machine learning techniques to the analysis of soil ecological data bases: Relationships between habitat features and Collembolan community characteristics. Soil Biol Biochem 32 (2): 197-209. DOI: 10.1016/S0038-0717(99)00147-9

- Kampichler C, Wieland R, Calmé S, Weissenberger H, Arriaga-Weiss S (2010) Classification in conservation biology: A comparison of five machine-learning methods. Ecological Informatics 5 (6): 441-450. DOI: 10.1016/j.ecoinf.2010.06.003
- Kasten E, McKinley P, Gage S (2010) Ensemble extraction for classification and detection of bird species. Ecological Informatics 5 (3): 153-166. DOI: 10.1016/j.ecoinf.2010.02.003
- Kastens T, Featherstone A (1996) Feedforward backpropagation neural networks in prediction of farmer risk preference. Am J Agric Econ 78 (2): 400-415. URL: http://www.jstor.org/stable/1243712
- Keerthi SS, Gilbert EG (2002) Convergence of a Generalized SMO Algorithm for SVM Classifier Design. Machine Learning 46: 351-360. DOI: 10.1023/a:1012431217818
- Keogh E, Mueen A (2011) Curse of Dimensionality. In: Sammut C, Webb G (Eds) Encyclopedia of Machine Learning. Springer, 257-258 pp. DOI: 10.1007/978-0-387-30164-8_192
- Knudby A, Brenning A, LeDrew E (2010a) New approaches to modelling fish–habitat relationships. Ecological Modelling 221 (3): 503-511. DOI: 10.1016/j.ecolmodel.2009.11.008
- Knudby A, LeDrew E, Brenning A (2010b) Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques. Remote Sensing of Environment 114 (6): 1230-1241. DOI: 10.1016/j.rse.2010.01.007
- Kobler A, Adamic M (2000) Identifying brown bear habitat by a combined GIS and machine learning method. Ecol Modell 135: 291-300. DOI: 10.1016/S0304-3800(00)00384-7
- Kohonen T (1989) Self-Organization and Associative Memory. Springer Series in Information Sciences, 312 pp. DOI: 10.1007/978-3-642-88163-3
- Kon M, Plaskota L (2000) Information complexity of neural networks. Neural Netw 13 (3): 365-375.
- Kotsiantis SB (2007) Supervised Machine Learning: A Review of Classification Techniques. Informatica 31: 249-268.
- Kotsiantis SB, Zaharakis ID, Pintelas PE (2006) Machine learning: a review of classification and combining techniques. Artificial Intelligence Review 26 (3): 159-190. DOI: 10.1007/s10462-007-9052-3
- Koza J (1992) Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, 836 pp.
- Krasnopolsky V (2009) Neural network applications to solve forward and inverse problems in atmospheric and oceanic satellite remote sensing. In: Haupt S, Pasini A, Marzban C (Eds) Artificial Intelligence Methods in the Environmental Sciences. Springer Netherlands, Amsterdam, 191-206 pp.
- Lakshmanan V (2009) Automated analysis of spatial grids. In: Haupt S, Pasini A, Marzban C (Eds) Artificial Intelligence Methods in the Environmental Sciences. 2009. Springer Netherlands, Amsterdam, 329-346 pp.
- Laplace PS (1986) Memoir on the Probability of the Causes of Events. Statistical Science 1 (3): 364-378. DOI: 10.1214/ss/1177013621
- Lawler J, White D, Neilson R, Blaustein A (2006) Predicting climate-induced range shifts: Model differences and model reliability. Glob Chang Biol 12 (8): 1568-1584.

- Lee J, Kwak I, Lee E, Kim K (2007) Classification of breeding bird communities along an urbanization gradient using an unsupervised artificial neural network. Ecological Modelling 203: 62-71. DOI: 10.1016/j.ecolmodel.2006.04.033
- Lees B (1996) Sampling strategies for machine learning using GIS. In: Goodchild M, Steyaert L, Parks B, Johnston C, Maidment D, Crane M, Glendinning S (Eds) GIS and Environmental Modeling: Progress and Issues. GIS World Inc., Fort Collins, 39-42 pp. [ISBN 978-0-470-23677-2].
- Lees B, Ritman K (1991) Decision-tree and rule-induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environments. Environmental Management 15 (6): 823-831. DOI: 10.1007/bf02394820
- Lek S, Guégan J (1999) Artificial neural networks as a tool in ecological modelling, an introduction. Ecological Modelling 120: 65-73. DOI: 10.1016/S0304-3800(99)00092-7
- Lek S, Guégan J (2000) Artificial Neuronal Networks: Application to Ecology and Evolution. Springer Verlag, Berlin, 262 pp. DOI: 10.1007/978-3-642-57030-8
- Lek S, Belaud A, Baran P, Dimopoulos I, Delacoste M (1996a) Role of some environmental variables in trout abundance models using neural networks. Aquat. Living Resour. 9 (1): 23-29. DOI: 10.1051/alr:1996004
- Lek S, Delacoste M, Baran P, Dimopoulos I, Lauga J, Aulagnier S (1996b) Application of neural networks to modelling non linear relationships in ecology. Ecol Modell 90: 39-52. DOI: 10.1016/0304-3800(95)00142-5
- Levine E, Kimes D, Sigillito V (1996) Classifying soil-structure using neural networks. Ecol Modell 92: 101-108.
- Li J, Heap A, Potter A, Daniell J (2011) Application of machine learning methods to spatial interpolation of environmental variables. Environmental Modelling & Software 26 (12): 1647-1659. DOI: 10.1016/j.envsoft.2011.07.004
- Lim T, Loh W, Shih Y (2000) A comparison of prediction accuracy, complexity, and training time of thirty three old and new classification algorithms. Machine Learning 40 (3): 203-228. DOI: 10.1023/a:1007608224229
- Loh W (2014) Fifty Years of Classification and Regression Trees. International Statistical Review 82 (3): 329-348. DOI: 10.1111/insr.12016
- Lorena A, Jacintho LO, Siqueira M, Giovanni RD, Lohmann L, de Carvalho AP, Yamamoto M (2011) Comparing machine learning classifiers in potential distribution modelling. Expert Systems with Applications 38 (5): 5268-5275. DOI: 10.1016/j.eswa.2010.10.031
- Lorenz E (1963) Deterministic non-periodic flow. J Atmos Sci 20: 130-141.
- MacLeod N (2007) Automated Taxon Identification in Systematics: Theory, Approaches and Applications. CRC Press, 368 pp. [ISBN 9780849382055]
- Maier H, Dandy G (2000) Neural networks for the prediction and forecasting of water resource variables: A review of modelling issues and applications. Environ Model Softw 15 (1): 101-124. DOI: 10.1016/S1364-8152(99)00007-9
- Manel S, Dias JM, Buckton ST, Ormerod SJ (1999) Alternative methods for predicting species distribution: an illustration with Himalayan river birds. J Appl Ecology 36 (5): 734-747. DOI: 10.1046/j.1365-2664.1999.00440.x
- Marzban C (2003) A neural network for post-processing model output: ARPS. Mon Weather Rev 131: 1103-1111.

- Mastrorillo S, Lek S, Dauba F, Belaud A (1997) The use of artificial neural networks to predict the presence of small-bodied fish in a river. Freshwater Biology 38 (2): 237-246. DOI: 10.1046/j.1365-2427.1997.00209.x
- McKay R (2001) Variants of genetic programming for species distribution modelling - fitness sharing, partial functions, population evaluation. Ecol Modell 146: 231-241. DOI: 10.1016/S0304-3800(01)00309-X
- McKenna JJ (2005) Application of neural networks to prediction of fish diversity and salmonid production in the Lake Ontario basin. Trans Am Fish Soc 134 (1): 28-43.
- Merow C, Smith M, Edwards T, Guisan A, McMahon S, Normand S, Thuiller W, Wüest R, Zimmermann N, Elith J (2014) What do we gain from simplicity versus complexity in species distribution models? Ecography 37 (12): 1267-1281. DOI: 10.1111/ecog.00845
- Meynard C, Quinn J (2007) Predicting species distributions: a critical comparison of the most common statistical models using artificial species. Journal of Biogeography 34 (8): 1455-1469. DOI: 10.1111/j.1365-2699.2007.01720.x
- Micci-Barreca D (2001) A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. ACM SIGKDD Explorations Newsletter 3 (1): 27. DOI: 10.1145/507533.507538
- Miller J, Franklin J (2002) Modelling distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecol Modell 157: 227-247. DOI: 10.1016/S0304-3800(02)00196-5
- Mitchell M (1998) An Introduction to Genetic Algorithms. MIT Press, Cambridge, 221 pp.
- Moeyersoms J, Martens D (2015) Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. Decision Support Systems 72: 72-81. DOI: 10.1016/j.dss.2015.02.007
- Moguerza JM, Muñoz A (2006) Support vector machines with applications. Statistical Science 21 (3): 322-336.
- Mountrakis G, Im J, Ogole C (2011) Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing 66 (3): 247-259. DOI: 10.1016/j.isprsjprs.2010.11.001
- Muggleton S (1995) Inverse entailment and progol. New Generation Computing 13: 245-286. DOI: 10.1007/bf03037227
- Mulligan A, Brown L (1998) Genetic algorithms for calibrating water quality models. J Environ Eng 124 (3): 202-211. DOI: 10.1061/(ASCE)0733-9372(1998)124:3(202)
- Muttil N, Lee JW (2005) Genetic programming for analysis and real-time prediction of coastal algal blooms. Ecological Modelling 189: 363-376. DOI: 10.1016/j.ecolmodel.2005.03.018
- Olden J, Lawler J, Poff NL (2008) Machine Learning Methods Without Tears: A Primer for Ecologists. The Quarterly Review of Biology 83 (2): 171-193. DOI: 10.1086/587826
- Omar S, Ngadi A, Jebur HH (2013) Machine learning techniques for anomaly detection: An overview. International Journal of Computer Applications 79 (2): 33-41.
- Özesmi S, Tan C, Özesmi U (2006) Methodological issues in building, training, and testing artificial neural networks in ecological applications. Ecological Modelling 195: 83-93. DOI: 10.1016/j.ecolmodel.2005.11.012
- Pal M (2005) Random forest classifier for remote sensing classification. International Journal of Remote Sensing 26 (1): 217-222. DOI: 10.1080/01431160412331269698

- Park Y, Chon T (2007) Biologically-inspired machine learning implemented to ecological informatics. Ecological Modelling 203: 1-7. DOI: 10.1016/j.ecolmodel.2006.05.039
- Parsons S, Jones G (2000) Acoustic identification of twelve species of echoloating bat by discriminant function analysis and artificial neural networks. J Exp Biol 203: 2641-2656.
- Pasini A (2009) Neural network modeling in climate change studies. In: Haupt S, Pasini A, Marzban C (Eds) *Artificial Intelligence Methods in the Environmental Sciences*. Springer Netherlands, Amsterdam, 235-254 pp.
- Peters J, Baets BD, Verhoest NC, Samson R, Degroeve S, Becker PD, Huybrechts W (2007) Random forests as a tool for ecohydrological distribution modelling. Ecological Modelling 207: 304-318. DOI: 10.1016/j.ecolmodel.2007.05.011
- Peterson AT, Ortega-Huerta M, Bartley J, Sánchez-Cordero V, Soberón J, Buddemeier R, Stockwell DB (2002) Future projections for Mexican faunas under global climate change scenarios. Nature 416 (6881): 626-629. DOI: 10.1038/416626a
- Pineda F (1987) Generalization of back-propagation to recurrent neural networks. Phys. Rev. Lett. 59 (19): 2229-2232. DOI: 10.1103/physrevlett.59.2229
- Pouteau R, Meyer J, Taputuarai R, Stoll B (2012) Support vector machines to map rare and endangered native plants in Pacific islands forests. Ecological Informatics 9: 37-46. DOI: 10.1016/j.ecoinf.2012.03.003
- Pradhan B (2010) Manifestation of an advanced fuzzy logic model coupled with Geo-information techniques to landslide susceptibility mapping and their comparison with logistic regression modelling. Environmental and Ecological Statistics 18 (3): 471-493. DOI: 10.1007/s10651-010-0147-7
- Prasad A, Iverson L, Liaw A (2006) Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. Ecosystems 9 (2): 181-199. DOI: 10.1007/s10021-005-0054-1
- Quinlan JR, Cameron-Jones RM (1995) Induction of logic programs: FOIL and related systems. New Generation Computing 13: 287-312. DOI: 10.1007/bf03037228
- Quintero E, Thessen A, Arias-Caballero P, Ayala-Orozco B (2014) A statistical assessment of population trends for data deficient Mexican amphibians. PeerJ 2: e703. DOI: 10.7717/peerj.703
- Rasmussen C, Williams C (2006) Gaussian Processes for Machine Learning. MIT Press, Cambridge, 248 pp. [ISBN 026218253X]
- Reby D, Joachim J, Lauga J, Lek S, Aulagnier S (1998) Individuality in the groans of fallow deer (Dama dama) bucks. J Zoology 245 (1): 79-84. DOI: 10.1111/j.1469-7998.1998.tb00074.x
- Recknagel F (1997) ANNA - artificial neural network model for predicting species abundance and succession of blue-green algae. Hydrobiologia 349: 47-57.
- Recknagel F (2001) Applications of machine learning to ecological modelling. Ecol Modell 146: 303-310. DOI: 10.1016/S0304-3800(01)00316-7
- Recknagel F, Bobbin J, Whigham P, Wilson H (2000) Multivariate time-series modelling of algal blooms in freshwater lakes by machine learning. In: Vanrolleghem P, Lessard P (Eds) Proceedings of the 5th International Symposium WATERMATEX'2000 on Systems Analysis and Computing in Water Quality Management.
- Recknagel F, Bobbin J, Whigham P, Wilson H (2002) Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes. J Hydroinformatics 4 (2): 125-133.

- Reeves C, Rowe JE (2002) Genetic Algorithms - Principles and Perspectives. Springer, 332 pp. DOI: 10.1007/b101880
- Ribic C, Ainley D (1997) The relationships of seabird assemblages to physical habitat features in Pacific equatorial waters during spring 1984–1991. ICES Journal of Marine Science 54 (4): 593-599. DOI: 10.1006/jmsc.1997.0244
- Rogan J, Franklin J, Stow D, Miller J, Woodcock C, Roberts D (2008) Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. Remote Sens Environ 112 (5): 2272-2283. DOI: 10.1016/j.rse.2007.10.004
- Rosa ID, Marques AT, Palminha G, Costa H, Mascarenhas M, Fonseca C, Bernardino J (2015) Classification success of six machine learning algorithms in radar ornithology. Ibis 158 (1): 28-42. DOI: 10.1111/ibi.12333
- Ruck B, Walley W, Hawkes H (1993) Biological classification of river water quality using neural networks. In: Rzevski G, Pastor J, Adey R (Eds) Applications of Artificial Intelligence in Engineering VIII. Elsevier Applied Science, 361-372 pp.
- Rumelhart D, Hinton G, Williams R (1986) Learning internal representations by error propagation. In: Rumelhart D, McClelland J, Group P (Eds) Parallel Distributed Processing. MIT Press, Cambridge, 318-362 pp.
- Saar-Tsechansky M, Provost F (2007) Handling Missing Values when Applying Classification Models . Journal of Machine Learning Research 8: 1625-1657.
- Sarkar S, Pressey R, Faith D, Margules C, Fuller T, Stoms D, Moffett A, Wilson K, Williams K, Williams P, Andelman S (2006) Biodiversity Conservation Planning Tools: Present Status and Challenges for the Future. Annual Review of Environment and Resources 31 (1): 123-159. DOI: 10.1146/annurev.energy.31.042606.085844
- Sastry K, Goldberg D, Kendall G (2013) Genetic Algorithms. Search Methodologies. Springer, 93-117 pp. DOI: 10.1007/978-1-4614-6940-7_4
- Scardi M (1996) Artificial neural networks as empirical models for estimating phytoplankton production. Marine Ecology Progress Series 139: 289-299.
- Scardi M, Harding LJ (1999) Developing an empirical model of phytoplankton primary production: a neural network case study. Ecological Modelling 120: 213-223.
- Schultz A, Wieland R, Lutze G (2000) Neural networks in agroecological modelling - stylish application or helpful tool? Computers and Electronics in Agriculture 29 (1): 73-97. DOI: 10.1016/S0168-1699(00)00137-X
- Seginer I, Boulard T, Bailey BJ (1994) Neural Network Models of the Greenhouse Climate. Journal of Agricultural Engineering Research 59 (3): 203-216. DOI: 10.1006/jaer.1994.1078
- Segurado P, Araújo M (2004) An evaluation of methods for modelling species distributions. Journal of Biogeography 31 (10): 1555-1568. DOI: 10.1111/j.1365-2699.2004.01076.x
- Semtner A (1995) Modeling Ocean Circulation. Science 269: 1379-1380.
- Siddique MN, Tokhi MO (2001) Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on. 4. IJCNN'01. IEEE, 2673-2678 pp. DOI: 10.1109/ijcnn.2001.938792
- Simmonds E, Armstrong F, Copland P (1996) Species identification using wideband backscatter with neural network and discriminant analysis. ICES J Mar Sci. 53: 189-195.
- Simonite T (2016) Algorithms That Learn with Less Data Could Expand AI's Power. MIT Technology Review 601551: e. URL: https://www.technologyreview.com/s/601551/algorithms-that-learn-with-less-data-could-expand-ais-power/

- Sokal RR, Rohlf FJ (2011) Biometry. W.H. Freeman, 937 pp. [ISBN 0716786044]
- Sosik H, Olson R (2007) Automated taxonomic classification of phytoplankton sampled with imaging in flow cytometry. Limnol Oceanogr Methods 5 (6): 204-216. DOI: 10.4319/lom.2007.5.204
- Spitz F, Lek S (1999) Environmental impact prediction using neural network modelling. An example in wildlife damage. J Appl Ecology 36 (2): 317-326. DOI: 10.1046/j.1365-2664.1999.00400.x
- Srinivasan A, King R, Muggleton S, Sternberg M (1997) Carcinogenesis prediction using inductive logic programming. In: Lavrač N, Keravnou E, Zupan B (Eds) Intelligent Data Analysis in Medicine and Pharmacology. Springer US, New York, 243-260 pp.
- Stockwell D (1999) Genetic Algorithms II. In: Fielding AH (Ed.) Machine Learning Methods for Ecological Applications. Springer US, New York, 123-144 pp.
- Stockwell D, Noble I (1992) Induction of sets of rules from animal distribution data: A robust and informative method of analysis. Math Comput Simul 33: 385-390. DOI: 10.1016/0378-4754(92)90126-2
- Stockwell D, Peters D (1999) The GARP modelling system: problems and solutions to automated spatial prediction. International Journal of Geographical Information Science 13 (2): 143-158. DOI: 10.1080/136588199241391
- Sutherland W, Pullin A, Dolman P, Knight T (2004) The need for evidence-based conservation. Trends in Ecology & Evolution 19 (6): 305-308. DOI: 10.1016/j.tree.2004.03.018
- Tan S, Smeins F (1996) Predicting grassland community changes with an artificial neural network model. Ecol Modell. Ecological Modelling 84: 91-97. DOI: 10.1016/0304-3800(94)00131-6
- Termansen M, McClean C, Preston C (2006) The use of genetic algorithms and Bayesian classification to model species distributions. Ecological Modelling 192: 410-424. DOI: 10.1016/j.ecolmodel.2005.07.009
- Thessen A, Patterson D (2011) Data issues in the life sciences. ZooKeys 150: 15-51. DOI: 10.3897/zookeys.150.1766
- Thuiller W (2003) BIOMOD - optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biol 9 (10): 1353-1362. DOI: 10.1046/j.1365-2486.2003.00666.x
- Tscherko D, Kandeler E, Bárdossy A (2007) Fuzzy classification of microbial biomass and enzyme activities in grassland soils. Soil Biology and Biochemistry 39 (7): 1799-1808. DOI: 10.1016/j.soilbio.2007.02.010
- Vander Zanden MJ, Olden J, Thorne J, Mandrak N (2004) Predicting occurrences and impacts of smallmouth bass introductions in north temperate lakes. Ecological Applications 14 (1): 132-148. DOI: 10.1890/02-5036
- Vayssières M, Plant R, Allen-Diaz B (2000) Classification trees: An alternative non-parametric approach for predicting species distributions. Journal of Vegetation Science 11 (5): 679-694. DOI: 10.2307/3236575
- Veropoulos K, Campbell C, Cristianini N (1999) Controlling the sensitivity of support vector machines. Proceedings of the International Joint Conference on Artificial Intelligence. IJCA1999.
- Wagstaff K (2012) Machine learning that matters. In: Langford J, Pineau J (Eds) Proceedings of the 29th International Conferene on Machine Learning. California Institute of Technology, 298-303 pp.

- Walley W, Džeroski S (1996) Biological monitoring: A comparison between Bayesian, neural and machine learning methods of water quality classification. In: Denzer R, Schimak G, Russell D (Eds) Environmental Software Systems: Proceedings of the International Symposium on Environmental Software Systems. Springer US, New York, 229-240 pp. DOI: 10.1007/978-0-387-34951-0_20
- Walley W, Hawkes H, Boyd M (1992) Application of Bayesian inference to river water quality surveillance. In: Grierson D, Rzevski G, Adey R (Eds) Applications of Artificial Intelligence in Engineering VII. Elsevier, 1030-1047 pp.
- Walley W, Martin R, O'Connor M (2000) Self-organising maps for the classification and diagnosis of river quality from biological and environmental data. In: Denzer R, Swayne D, Purvis M, Schimak G (Eds) Environmental Software Systems: Environmental Information and Decision Support. Springer US, New York, 27-41 pp.
- Wang P, Ruan D, Kerre E (2007) Fuzzy Logic. Springer, 459 pp. DOI: 10.1007/978-3-540-71258-9
- Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Applied and Environmental Microbiology 73 (16): 5261-5267. DOI: 10.1128/aem.00062-07
- Webb T, Vanden Berghe E, O'Dor R (2010) Biodiversity's Big Wet Secret: The Global Distribution of Marine Biological Records Reveals Chronic Under-Exploration of the Deep Pelagic Ocean. PLoS ONE 5 (8): e10223. DOI: 10.1371/journal.pone.0010223
- Wieland R (2008) Fuzzy models. In: Jørgensen S, Fath B (Eds) Encyclopedia of Ecology. Elsevier, Amsterdam, 1717-1726 pp.
- Wieland R, Mirschel W (2008) Adaptive fuzzy modeling versus artificial neural networks. Environmental Modelling & Software 23 (2): 215-224. DOI: 10.1016/j.envsoft.2007.06.004
- Wiley EO, McNyset K, Peterson T, Robins R, Stewart A (2003) Niche Modeling Perspective on Geographic Range Predictions in the Marine Environment Using a Machine-learning Algorithm. Oceanography 16 (3): 120-127. DOI: 10.5670/oceanog.2003.42
- Williams J, Kessinger C, Abernathy J, Ellis S (2009) Fuzzy logic applications. In: Haupt S, Pasini A, Marzban C (Eds) Artificial Intelligence Methods in the Environmental Sciences. Springer Netherlands, Amsterdam, 347-378 pp.
- Worner SP, Gevrey M (2006) Modelling global insect pest species assemblages to determine risk of invasion. Journal of Applied Ecology 43 (5): 858-867. DOI: 10.1111/j.1365-2664.2006.01202.x
- Wu A, Hsieh W, Tang B (2006) Neural networks forecasts of the tropical Pacific sea surface temperatures. Neural Networks 19: 145-154.
- Yen GG, Lu H (2000) Combinations of Evolutionary Computation and Neural Networks, 2000 IEEE Symposium on. San Antonio, Texas, USA. IEEE, 168-175 pp. DOI: 10.1109/ecnn.2000.886232
- Young G (2009) Implementing a neural network emulation of a satellite retrieval algorithm. In: Haupt S, Pasini A, Marzban C (Eds) Artificial Intelligence Methods in the Environmental Sciences. Springer Netherlands, Amsterdam, 207-216 pp.
- Zhao K, Popescu S, Zhang X (2008) Bayesian Learning with Gaussian Processes for Supervised Classification of Hyperspectral Data. Photogrammetric Engineering & Remote Sensing 74 (10): 1223-1234. DOI: 10.14358/pers.74.10.1223

- Zhao K, Popescu S, Meng X, Pang Y, Agca M (2011) Characterizing forest canopy structure with lidar composite metrics and machine learning. Remote Sensing of Environment 115 (8): 1978-1996. DOI: 10.1016/j.rse.2011.04.001